Transactions of the Philological Society Volume 103:2 (2005) 121–146

VERTICAL AND HORIZONTAL TRANSMISSION IN LANGUAGE EVOLUTION[1]

By William S-Y. Wang[a] and James W. Minett[b]
[a]*Chinese University of Hong Kong and Academia Sinica, Taiwan,* [b]*Chinese University of Hong Kong*

Abstract

It has been observed that borrowing within a group of genetically related languages often causes the lexical similarities among them to be skewed. Consequently, it has been proposed that borrowing can sometimes be inferred from such skewing. However, heterogeneity in the rate of lexical replacement, as well as borrowing from other languages, can also give rise to skewed lexical similarities. It is important, therefore, to determine to what degree skewing is a statistically significant indicator of borrowing. Here, we describe a statistical hypothesis test for detecting language contact based on skewing of linguistic characters of arbitrary type. Significant probabilities of correct detection of contact are maintained for various contact scenarios, with low false alarm probability. Our experiments show that the test is fairly robust to substantial heterogeneity in the retention rate, both across characters and across lineages, suggesting that the method can provide an objective criterion against which claims of significant skewing due to contact can be tested, pointing the way for more detailed analysis.

[1]This work has been supported in part by grants 1224/02H and 1127/04H awarded by the Research Grants Council of Hong Kong to the Chinese University of Hong Kong. We would like to thank the anonymous reviewers of an earlier draft of this paper for their constructive criticism. We are also grateful to Drs. Jinyun Ke and Feng Wang, and Miss Joyce Cheung for their helpful comments. Additional supplementary material can be found at http://www.philsoc.org.uk/transactions.asp

1. Introduction

The tree diagram allows the hierarchy of language splits that are hypothesized to have taken place in a language stock to be displayed clearly and simply. However, the underlying assumption that languages split discretely into two (or more) lineages, each evolving independently thereafter by vertical transmission only, is far from realistic. As Schmidt recognized when he proposed the *Wellentheorie*, languages often do not split nearly so cleanly as supposed in the *Stammbaumtheorie*. Furthermore, innovations that arise in one language may come to be acquired by other nearby languages with which it comes into contact. When such horizontally transmitted innovations are incorrectly interpreted as vertically transmitted innovations, and are used to infer a language tree, the topology of the tree provides a warped representation of the genetic relationships. In order to prepare tree diagrams that accurately reflect only the genetic relationships among a set of languages, the two main mechanisms by which innovations come to be shared must be distinguished, and the horizontal transmission filtered out. Alternatively, if a hybrid picture of the evolution of a set of languages is sought, comprising both modes of transmission, the tree diagram must be discarded and replaced by a network diagram on which are marked both the lines of descent by vertical transmission and the contact events involving horizontal transmission.

One method of distinguishing vertical and horizontal transmission, which can be traced back to Hübschmann's (1875) work on Armenian, is to stratify the correspondences between two lineages and to recognize that only the oldest stratum can possibly reflect the vertically transmitted signal. This idea has been taken up by Wen (1940), and more recently by Sagart & Xu (2001), to resolve contact among Sino-Tibetan languages, and forms part of the Distillation Procedure for reconstruction recently proposed by Wang (2004).

Other, more quantitative approaches to detecting horizontal transmission have also been proposed. One such class of methods is based on cladistics. For example, Warnow and Ringe, and their colleagues, have promoted the use of cladistic methods in linguistic classification, publishing a number of papers (e.g., Warnow *et al.* 1995; Ringe *et al.* 2002) in which they apply their own implementation of the maximum compatibility method to refine the classification of Indo-European. In their approach, determination of the optimal position of each sub-group is undertaken by seeking the topologies — there might be more than one — that are compatible with the greatest number of characters. The remaining, incompatible characters are viewed as having been subject to non-genetic processes, such as borrowing, and are not used to determine the optimal topologies. Application of the computational technique *answer set programming* to the automatic assessment of contact-induced innovations that are optimal within the Warnow-Ringe paradigm is currently under investigation (Erdem *et al.* 2003; Brooks *et al.* 2005), and shows promising results.

Minett & Wang (2003) have described another cladistic method for detecting borrowed characters. In this method, the optimal trees are sought according to the maximum parsimony criterion. Each innovation is assumed to arise independently only once, hence instances of some innovation after its first occurrence are considered to be due to contact. The method has been applied to detect lexical borrowing among representative dialects of each of the seven main sub-groups of Chinese. More tests must be performed to check that these cladistic methods are indeed detecting contact-induced change and not leading toward plausible, but spurious, inferences.

Methods based on lexicostatistics have also been attempted. A number of Africanists, including Hinnebusch (1999) and Heine (1971) before him, have noted that the lexical similarities between sub-families of closely genetically related languages tend to be approximately equal when only genetic effects have influenced the languages, but tend to be different when borrowing or other non-genetic effects have influenced them. This observation has led Hinnebusch (1996) to propose a lexicostatistical concept for identifying languages that have come into contact by looking for skewing in the lexical similarities among them — s*kewing* is just the difference that is observed between the similarities of one language with respect to two other languages; a more formal definition of skewing is given in Section 2.

Hinnebusch cites a comment by Heine (1974: 17), regarding possible contact among three Nilotic languages, which we repeat here to illustrate the concept (Hinnebusch, 1999: 177):

> The Nilotic languages Samburu and Nandi share 9.9 percent lexical resemblances on the basis of the 200-word list. The percentage between Masai and Nandi, on the other hand, amounts to 15.7. These two languages have been in close contact over the last few centuries. It seems reasonable to assume that the difference of 5.8 percent between Samburu/Nandi and Masai/Nandi is a result of the process of borrowing which took place between Masai and Nandi.

The claim that the 5.8% difference, or *skewing*, between the lexical similarities observed for Samburu/Nandi and Masai/Nandi is due to borrowing between Masai and Nandi is certainly intuitively appealing. However, notwithstanding the fact that Masai and Nandi are known to have come into contact recently, an alternative explanation for the skewing noted by Heine might simply be that Masai has been more conservative than Samburu since splitting from it, thereby causing the lexical similarity between Masai/Nandi to exceed that between Samburu/Nandi. Another possible explanation is that Samburu might have come into contact with some other language, causing some lexical items that had hitherto been cognate with lexical items in Nandi to be replaced, so reducing the lexical similarity for Samburu/Nandi.

Chen (2000) has proposed a lexicostatistical method for identifying whether a pair of languages have come into contact. The method, called *rank analysis*, divides the lexical items into groups, or *ranks*, that are known (or assumed) to have different average rates of replacement, and works with the lexical similarities between the languages in each of those ranks. In its simplest implementation, *universal rank analysis*, the Swadesh basic words are grouped into two ranks: Rank 1, consisting of the Swadesh 100 word-list (Swadesh, 1955), and Rank 2, the remaining words of the Swadesh 200 word-list (Swadesh, 1951). The lexical similarity between the pair of languages is then calculated for each rank. Based on the assumption that the Rank 1 words are more conservative and less susceptible to borrowing than the Rank 2 words, the two languages are inferred to be genetically related only if the Rank 1 similarity exceeds the Rank 2 similarity; otherwise, the shared resemblances are inferred to have come about through contact. Used in conjunction with stratification, whereby the rank analysis is applied only to the oldest detected stratum, the method may prove to be a powerful tool for detecting horizontal transmission. However, it is not yet clear how accurate this method is.

Several other lexicostatistical methods for detecting horizontal transmission have also been developed: For example, both Sankoff (1972) and Embleton (1981; 1986) have extended the traditional implementation of lexicostatistics to account for both heterogeneous retention rates and borrowing, modelling the borrowing between languages in terms of their geographic neighbourhood. However, these methods do not in themselves allow the detection of horizontal transmission. Rather, they use estimates of the borrowing rates between neighbouring languages to improve the classification produced by the lexicostatistical analyses.

In his study of the settlement of Taiwan, Wang (1989) postulated that patterns in the lexicostatistical error matrix (the absolute difference between the input lexical distances and the distances reconstructed from the optimal lexicostatistical tree) are indicative of horizontal transmission. While this approach has produced suggestive results, its efficacy is yet to be verified.

In addition to developing a cladistic method for detecting contact, Minett & Wang (2003) also proposed a lexicostatistical approach to detecting contact, hypothesizing that branches of negative length in the trees built by distance-based tree-building algorithms, such as Neighbor-Joining (Saitou and Nei, 1987), are indicative of contact. This hypothesis, however, turned out to be false.

Yet another approach, suggested by Cavalli-Sforza *et al.* (1994) for detecting admixture among human populations, is to use bootstrapping in conjunction with an arbitrary tree-building algorithm. Bootstrapping works by generating multiple trees, with one or more randomly selected languages removed from the analysis each time. The stability of the topologies so produced are then examined — clusters of languages that are grouped together for many bootstrap samples are considered to be representative of valid genetic relationships. But when a language is unstable in the sampling, shifting from one sub-grouping to another, it is considered likely that that language has come into contact with other languages. The multiple allegiances of such a language may perhaps be identified by examining for which bootstrap samples it shifts sub-group. This method has been applied to Indo-European by Ogura and Wang (1996) with some success, correctly detecting the heavy lexical borrowing by English from both French and the Scandinavian languages.

It is also important to mention the split decomposition method for phylogenetic analysis (Bandelt & Dress, 1992). The majority of classification algorithms that have been applied to historical linguistics constrain the topologies that are produced to be trees. However, as we have mentioned, horizontal transmission cannot be shown on a language tree, and actually warps the tree away from representing genetic relationships accurately. Vertically transmitted characters and horizontally transmitted characters, if not distinguished, tend to produce contradictory sub-groupings on the tree — in other words, they tend to be incompatible. The split decomposition method, however, does not constrain the topology to be a tree, but transforms the characters into a set of splits to construct a so-called *splits graph*. Only when the characters are compatible — suggesting that there has been no horizontal transmission — does the split decomposition method construct a tree; otherwise, a tree-like network is constructed. As more characters become subject to horizontal transmission, so the departure from a tree topology becomes more pronounced.

Our aim in this paper is to place Hinnebusch's idea for using skewing to detect language contact on firm ground by deriving a statistical hypothesis test that can detect contact under idealized conditions at prescribed levels of significance, and to investigate its level of performance under several less-idealized conditions. The paper is laid out as follows. Section 2 summarizes the skewing concept and our implementation of Hinnebusch's method for detecting contact among an arbitrary number of genetically related languages. In Section 3, we show results for a number of contact scenarios that illustrate the robustness of the test. Some concluding remarks are given in Section 4.

2. The skewing method for detecting language contact

The skewing method for inferring language contact outlined by Hinnebusch is a similarity-based lexicostatistical method. In a standard lexicostatistical approach, the lexical similarity of two languages is calculated by counting the proportion of some set of pre-selected meanings for which the corresponding glosses appear to be reflexes of the same etymon. Languages having a greater lexical similarity are considered to be more closely genetically related than languages having a lesser lexical similarity.

There are a number of theoretical problems with the lexicostatistical method: Sometimes, no word can found to express a particular meaning. At other times, multiple words are found to correspond to a certain meaning — which word should the linguist use to encode the character? However, looking at these problems from the viewpoint of statistics, the lexical similarity calculated for any two languages is simply an *estimate* of their similarity based on noisy data. As long as there are not too many such noisy characters and the linguist adopts a consistent approach to handling them, we believe that the lexical similarity can still be a useful tool for estimating how closely related are two languages. Also, Blust (2000) has reminded us that use of lexicostatistics can lead to incorrect classifications when the retention rate across lineages is heterogeneous. While it is true that lexicostatistics can perform poorly, even for only slight heterogeneity in the retention rate, it remains an important question as to *when* and *how often* lexicostatistics performs poorly. We emphasise, however, that in the skewing method described here, no attempt is made to actually classify languages using lexicostatistics.

2.1. Skewing

In order to explain more clearly what we mean by skewing and how skewing might be used to detect language contact, we find it convenient to make use of certain concepts used in cladistics. In cladistics, a *character* can be defined, rather loosely, as some feature of the taxa being classified, here languages, that allows them to be categorized on the basis of the different *character states* that selected characters manifest. Suppose that character state data is available for two sets of languages that are known to be members of two distinct, genetically related sub-groups of some language family, but for which the genetic relationships within each of the two sub-groups are unknown. Our aim is to formalize Hinnebusch's method into a statistical hypothesis test that can be used to infer whether there has been contact between the languages in two such sub-groups.

We begin by defining the *skewing* between two sibling languages, A and B, with respect to a third language, C, as the similarity of A and C minus the similarity of B and C. The similarity measure can be simply the usual lexical similarity that is adopted in lexicostatistical studies (e.g., as in Dyen *et al.* 1992) or some other measure of similarity based on characters of arbitrary type. If contact has occurred between, say, recipient language A and donor language C, A will tend to have a higher similarity with C than does B, resulting in positive skewing between A and B with respect to C. This tendency for contact to induce positive skewing forms the basis of the skewing method. As Hinnebusch (1999: 184) observes, "languages which group together lexicostatistically will tend to have a numerical symmetry with other noncontiguous languages in the comparison set if in fact the grouped languages form a genetic group." In other words, we would expect languages within one sub-group of languages to exhibit little skewing with respect to languages of another, related sub-group of languages. When, nonetheless, skewing is observed, language contact is one possible cause. We also define the *aggregate skewing* of a language, A, with respect to another language, C, as the average skewing between A and each of its siblings with respect to C.

The potential use of skewing as an indicator of language contact is best explained by means of an example. Consider the lexical similarities shown in Table 1 among eight Bantu dialects: four Mijikenda dialects (Chonyi, Giriyama, Duruma and Digo) and four Comorian dialects (Ngazija, Mwali, Nzuani and Maore), all members of the Sabaki sub-group of Bantu (data from Hinnebusch, 1999).

Calculation of the aggregate skewing for the Mijikenda dialect Digo with respect to each of the Comorian dialects is summarized

Table 1. Lexical similarities among two sub-groups of Nilotic languages, after Hinnebusch (1999).

Lexical Similarity (%)	Mijikenda: Chonyi	Giriyama	Duruma	Digo	Comorian: Ngazija	Mwali	Nzuani	Maore
Chonyi	100	81	78	68	59	60	59	59
Giriyama		100	77	66	60	58	59	60
Duruma			100	70	60	58	58	59
Digo				100	56	54	56	59
Ngazija					100	81	77	80
Mwali						100	83	84
Nzuani							100	83
Maore								100

in Table 2. So, for example, the aggregate skewing of Digo with respect to Ngazija is $-3\frac{2}{3}$ per cent.

Proceeding in this way, we can calculate the aggregate skewing for each of the Mijikenda dialects with respect to each of the Comorian dialects and *vice versa*, the results for which are shown in Table 3.

Notice that the aggregate skewing is not symmetric. For example, the skewing of Digo with respect to Ngazija, $-3\frac{2}{3}$ per cent, does not equal that of Ngazija with respect to Digo, $-1/3$ per cent. This is because the former value is obtained by summing the skewing for Digo and Ngazija over all the Mijikenda dialects while the latter value is obtained by summing over all the Comorian dialects. These values indicate that Digo is lexically less similar to Ngazija than are the other Mijikenda dialects, but that Ngazija is about as similar to

Table 2. An example of the calculation of aggregate skewing. Aggregate skewing is calculated for the Mijikenda dialect Digo with respect to four Comorian dialects.

(%)	Comorian: Ngazija	Mwali	Nzuani	Maore
Chonyi	56–59 = −3	54–60 = −6	56–59 = −3	59–59 = 0
Giriyama	56–60 = −4	54–58 = −4	56–59 = −3	59–60 = −1
Duruma	56–60 = −4	54–58 = −4	56–58 = −2	59–59 = 0
$\delta S_{\text{Digo}}^{\text{Comorian}}$	$-3\frac{2}{3}$	$-4\frac{2}{3}$	$-2\frac{2}{3}$	$-1/3$

Table 3(a). Aggregate skewing percentages of the Mijikenda dialects with respect to the Comorian dialects.

$\delta S^{\text{Comorian}}_{\text{Mijikenda}}$	Ngazija	Mwali	Nzuani	Maore
Chonyi	+1/3	$+3\frac{1}{3}$	$+1\frac{1}{3}$	−1/3
Giriyama	$+1\frac{2}{3}$	+2/3	$+1\frac{1}{3}$	+1
Duruma	$+1\frac{2}{3}$	+2/3	0	−1/3
Digo	$-3\frac{2}{3}$	$-4\frac{2}{3}$	$-2\frac{2}{3}$	−1/3

Table 3(b). Aggregate skewing of the Comorian dialects with respect to the Mijikenda dialects.

$\delta S^{\text{Mijikenda}}_{\text{Comorian}}$	Chonyi	Giriyama	Duruma	Digo
Ngazija	−1/3	+1	$+1\frac{2}{3}$	−1/3
Mwali	+1	$-1\frac{2}{3}$	−1	−3
Nzuani	−1/3	−1/3	−1	−1/3
Maore	−1/3	+1	+1/3	$+3\frac{2}{3}$

Digo as are the other Comorian dialects. Digo is also negatively skewed with respect to both Mwali and Nzuani. One possible cause is that Digo has borrowed from a separate sub-group. Indeed, Hinnebusch suggests that this is so, arguing that Digo has come into heavy contact with Swahili. If this is indeed the case, negative skewing would seem to indicate contact with some language outside the group. However, an alternative explanation — one based only on the skewing data — is simply that Digo is less conservative than the other Mijikenda dialects, causing it to exhibit fewer similarities with the Comorian dialects than its Mijikenda siblings. This would also account for the relatively low similarity of Digo with its siblings (shown in Table 1).

Examining Table 3 we see large magnitude positive skewing for Maore with respect to Digo, $+3\frac{2}{3}$ per cent, and for Chonyi with respect to Mwali, $+3\frac{1}{3}$ per cent. If then contact between two languages tends to induce positive skewing, these values imply possible contact between Maore and Digo, and between Chonyi and Mwali. But what is the probable direction of transmission? Consider the case of Chonyi and Mwali. If the direction of transmission were from Mwali to Chonyi, we would expect some of

Table 4. Parameters values used to generate pseudo-random character state data in the absence of contact.

Number of languages in sub-group Λ_1:	5
Number of languages in sub-group Λ_2:	5
Time depth of family:	2
Time depth of sub-groups:	1
Retention rate:	90%
Number of characters:	100

the character states acquired from Mwali by borrowing to also be present in the other Comorian dialects. Therefore, in addition to positive skewing of Chonyi with respect to Mwali, we would also expect to observe at least some positive skewing of Chonyi with respect to the other Comorian dialects. On the other hand, if the direction of transmission were from Chonyi to Mwali, we would not expect the lexical similarity between Chonyi and Mwali's Comorian siblings to be affected. Consequently, we would not expect any significant positive skewing of Chonyi with respect to the other Comorian dialects. Examining Table 3, it is apparent that horizontal transmission from Mwali into Chonyi is the more probable scenario. But of course, the skewing might be caused simply by heterogeneity of the retention rates.

2.2. Distribution of Aggregate Skewing — no contact

We proceed by deriving the distribution of aggregate skewing when there is no contact between the two sub-groups. Pseudo-random character state data are generated for ten languages that have split into two sub-groups, each sub-group comprising five languages.[2] The parameter values chosen for this experiment are summarized in Table 4; there is no contact. Note that the retention rate is set to be homogeneous — both across lineages and across characters — at 90%. We estimate the distribution of aggregate skewing by Monte-Carlo simulation, observing the relative frequency of different

[2]The algorithm used to generate the pseudo-random character state data is described in Appendix A in the supplementary material, available at http://ling.man.ac.uk/More/philsoc/Transactions/html.

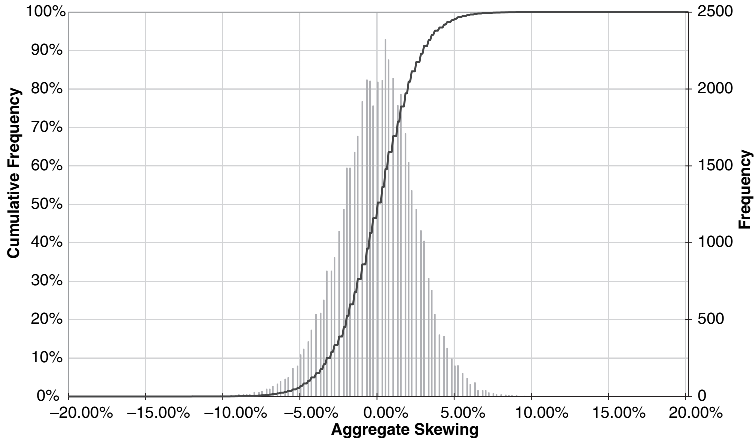


Figure 1. Distribution of aggregate skewing when there is no language contact. The mean value of aggregate skewing is zero.

values of aggregate skewing in multiple runs of the algorithm. Figure 1 shows the distribution of aggregate skewing for all pairs of languages observed over 1000 runs with the parameters specified in Table 4 — both the frequency (the vertical bars) and the cumulative frequency (the curve) of aggregate skewing are shown in the figure. Notice that the aggregate skewing is approximately Gaussian distributed with zero mean.

2.3. Distribution of Aggregate Skewing — contact

We now examine how the distribution of skewing changes when there is contact between the two sub-groups by injecting borrowing of various degrees between a single donor language and a single recipient language, one language in each sub-group. For the first such set of runs, the parameter values of the algorithm are set to those values specified in Table 4; the contact rate is set to 10%.

Figures 2 & 3 summarize the results of this experiment: Figure 2 shows the distribution of aggregate skewing for the pairs of languages that have not come into contact; Figure 3 shows the distribution of aggregate skewing of the recipient language with

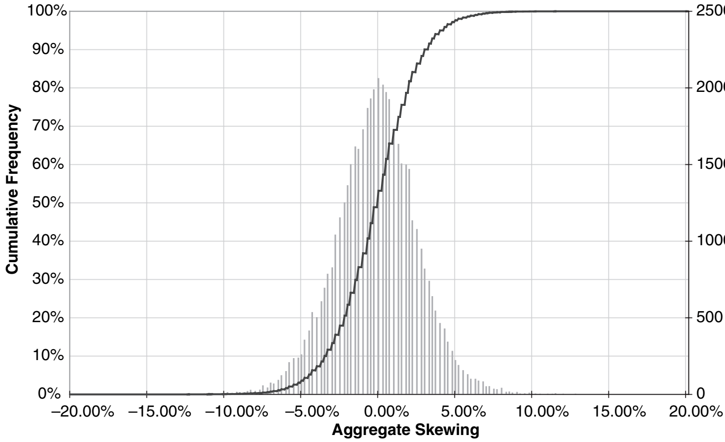


Figure 2. Distribution of aggregate skewing for languages that have not come into contact. The mean value of aggregate skewing is approximately zero (–0.1%).

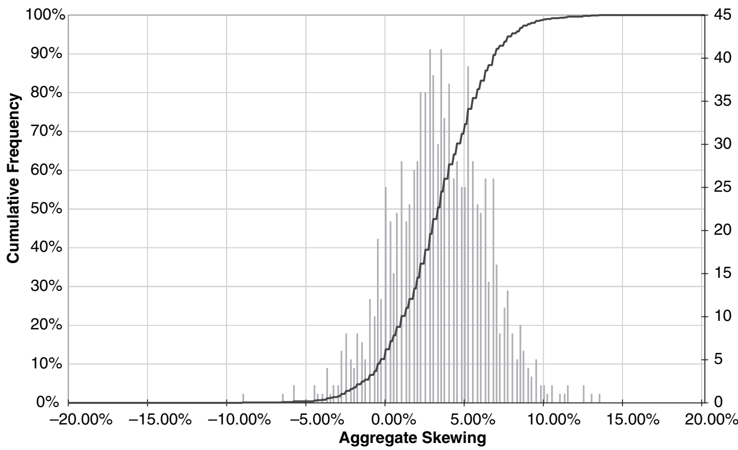


Figure 3. Distribution of aggregate skewing for languages that have come into contact (recipient with respect to donor). The mean value of aggregate skewing significantly exceeds zero (+3.3%).

respect to the donor language. In both cases, the distribution of aggregate skewing is roughly Gaussian. For the pairs of languages that have not come into contact, the mean level of aggregate skewing is −0.1% (Figure 2), only slightly lower than the zero-mean skewing observed under the no-contact scenario (cf. Figure 1). However, due to the contact between the donor and recipient languages, we expect these two languages to exhibit positive aggregate skewing. In fact, the observed mean level of aggregate skewing of the recipient language with respect to the donor language is +3.3% (Figure 3). We find then that language contact tends to induce positive aggregate skewing between the donor and recipient, but does not greatly affect the aggregate skewing for languages that have not come into contact.

2.4. Hypothesis test for detecting language contact

The above findings point to a method for identifying languages that have come into contact — positive aggregate skewing tends to indicate language contact. But what amount of aggregate skewing is a significant indicator of contact? As Figure 1 indicates, when there is no contact, less than 5% of language pairs have aggregate skewing of +3.7% or greater; but slightly more than 5% of language pairs have aggregate skewing of +3.6% or greater. Hence, for a 5% probability of false alarm (the *significance*), language contact can be inferred whenever the aggregate skewing exceeds the threshold value, θ, of +3.7%. Looking back at Figure 2, which shows the distribution of aggregate skewing among language pairs which have not come into contact when contact has occurred between some other language pair, we observe that 6.0% of language pairs exhibit significant skewing that exceeds the threshold, only slightly greater than the specified level of significance of 5%. This is further evidence (for this set of parameters at least) that contact between a single pair of languages does not induce significant aggregate skewing between other language pairs.

From Figure 3, we observe significant aggregate skewing of the recipient language with respect to the donor language with frequency 42.2%. Thus we can correctly infer contact between a single pair of languages with probability roughly 42% as long as we are prepared to accept ~6% chance of incorrectly inferring contact between a pair of languages between which no contact has occurred.

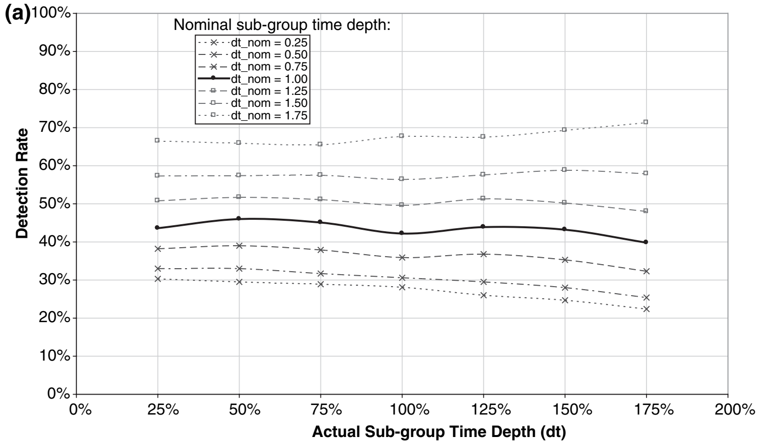
3. Results and discussion

We now examine the performance of the method described in Section 2 for inferring contact in a variety of contact situations. One thousand pseudo-random data sets are generated for each experiment detailed below. In each case, the observed distribution of aggregate skewing is determined so that the probabilities of incorrectly inferring contact, the *false alarm rate*, and of correctly inferring contact, the *detection rate*, can be estimated.

Experiment 1: The examples given in Section 2 were implemented based on the assumption that the sub-group time depth, as well as the time depth of the entire family, were assumed to be known, allowing an appropriate threshold value to be calculated. Here we consider the effect of setting the threshold based on an incorrect evaluation of the sub-group time depth. The sub-group time depth is set successively to 0.25, 0.50, …, 1.75, while the contact rate is set to 10%. However, for each value of the *actual* sub-group time depth, the performance is assessed for a threshold optimised for each value of the assumed, or *nominal*, sub-group time depth. Thus, in most cases, the chosen threshold is not optimised for the actual value of the sub-group time depth.

Figure 4 shows the resultant performance for 10% contact: (a) the detection rate, and (b) the false alarm rate. Examining the detection rate curves, we observe that for each value of the nominal sub-group time depth the detection rate is roughly constant for all actual values of the sub-group time-depth (although there is a slight dependence on the actual time depth for the more extreme values of the nominal time depth). This means that a reasonably accurate assessment of the detection rate can be obtained based on just the nominal sub-group time depth — the greater its value, the greater the detection rate.

Examining now the false alarm rate curves in Figure 4(b), we see that the false alarm rate certainly does depend on the values of both the actual sub-group time depth and the nominal sub-group time



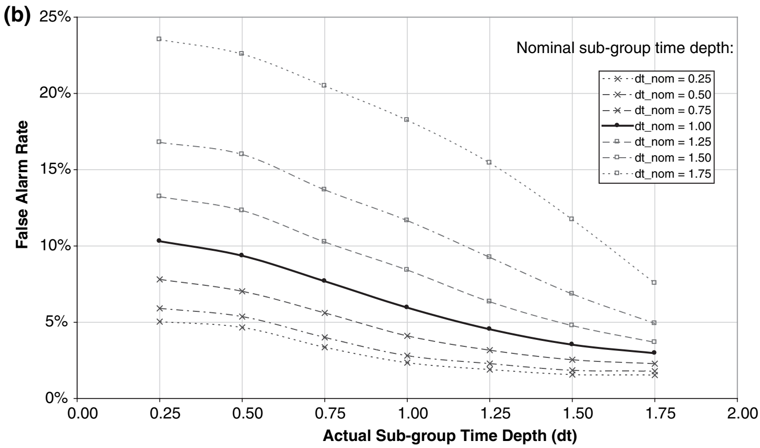


Figure 4(a). Detection rate for incorrect setting of the sub-group time depth. Performance is shown as a function of *actual* sub-group time depth (dt) for various values of the *nominal* sub-group time depth (dt_nom).
Figure 4(b). False alarm rate for incorrect setting of the sub-group time depth. Performance is shown as a function of *actual* sub-group time depth (dt) for various values of the *nominal* sub-group time depth (dt_nom).

depth. Since the former value is usually unknown, one can determine a robust threshold value by selecting the greatest nominal sub-group time depth for which the false alarm rate is no greater than some predetermined percentage and using the associated threshold. For example, if we are willing to accept at most a 10% probability of incorrectly inferring contact between languages between which no contact has taken place, we should select a threshold based on a nominal sub-group time depth of about 1.00, for which the false alarm rate for all values of the actual sub-group time depth is less than 10%. Since the detection rate is approximately constant for all values of the actual sub-group time depth, we can then read off the expected detection rate, about 40–45%, from Figure 4(a). If the sub-group time depth were actually 1.25, say, we would maintain a detection rate of about 44%, while the false alarm rate would be about 4%. Having some knowledge of the probable range of values of the sub-group time depth allows a tighter bound on the threshold to be set, thereby potentially achieving a lower false alarm rate and a higher detection rate.

For example, if we were seeking to detect contact between two sub-groups of Indo-European languages, we might assume the time depth of proto-Indo-European to be about 6 millennia and the sub-group time depth to be somewhat less, perhaps several millennia, depending on the particular sub-groups selected. On the contrary, if we were seeking to detect contact among the Northern and Southern dialects of Chinese,[3] we might pick the time depth of the proto-language of the family to be commensurate with that of Old Chinese, perhaps 2½ millennia, and the time depth of the sub-groups to be no more than 2 millennia. All other things being equal, we would therefore expect to observe significantly greater skewing among the Indo-European languages than among the Chinese dialects, even if there were no contact, because the Indo-European languages have had greater time to differentiate. Indeed, Wang (1997) finds the differentiation among seven representative Chinese

[3]The seven main dialects of Chinese are traditionally grouped into two first-order sub-groups: the Northern dialects, Mandarin, Xiang, Gan, Wu and Yue; and the Southern dialects, Min and Hakka. Nevertheless, the first-order sub-grouping of the Chinese dialects remains an open question and continues to be the subject of much research.

dialects to be roughly equal to that among the Romance languages of Indo-European based on lexical data collected by Xu (1991) and Dyen *et al* (1992). As a result, the threshold values for these two language families would likely be quite different.

Similar tests both for 20% contact rate and for a family time depth of 4.0, for example, reveal the same qualitative behaviour.

Experiment 2: So far, we have considered only homogeneous retention rates. However, we require the contact detection method described here to perform well also when the retention rates across lineages are heterogeneous. In this experiment, the retention rate across lineages is uniformly distributed on some range; the retention rate across characters is constant. The ranges investigated here are 90% ± 5%, 90% ± 10%, 85% ± 5% and 85% ± 10%, the results for which are compared to those for a retention rate of precisely 90%. The contact rate is set to 10% while all other parameters are set to the values shown previously in Table 4. In each case, the threshold is set for a nominal retention rate of 90%.

Table 5 summarizes the results of this experiment. The first point to note is that heterogeneity in the retention rate tends to cause both the detection rate and the false alarm rate to rise. For example, when the retention rate varies uniformly between the bounds 90% ± 10%, the detection rate increases by about 4½% while the false alarm rate increases by about 8%. While the increase in the detection rate is desirable, the increase in the false alarm rate is not.

Curiously, while the false alarm rate increases with mean retention rate for ±5% heterogeneity, it tends to decrease with mean retention rate for ±10% heterogeneity. This qualitative

Table 5. Detection rate and false alarm rate as a function of retention rate — retention rate heterogeneous across lineages (90% nominal retention rate; 10% contact).

Retention Rate:	90% (nominal)	90% ± 5%	90% ± 10%	85% ± 5%	85% ± 10%	80% ± 5%	80% ± 10%
Detection Rate (%)	42.2	41.3	46.0	61.3	59.5	72.3	67.5
False Alarm Rate (%)	6.0	8.7	14.4	9.5	14.6	10.1	13.2

behaviour is also observed for other sets of parameter values, suggesting that the method is quite robust when the degree of heterogeneity is not too great, in this case not exceeding ±10%. The results also indicate that the false alarm rate tends to deteriorate as the mean retention rate decreases. Control of the false alarm rate can be maintained by setting the threshold based on a suitably low nominal retention rate, but this causes a corresponding decrease in the detection rate. For example, setting the threshold based on a nominal retention rate of 80%, rather than 90% as in Table 5, produces the performance given in Table 6. For retention rate 85% ± 10%, the false alarm rate is reduced from 14.6% to 10.4% while the detection rate is reduced from 59.5% to 51.9%.

The final choice of the threshold value is a decision that must be made based on the investigator's aims: if a low false alarm rate is required, a relatively high threshold should be set; but if the aim is merely to determine a set of putative contact hypotheses to be studied by other methods in more detail, a lower, less conservative threshold can be set.

Experiment 3: Repeating Experiment 2, but with heterogeneity in the retention rate across characters rather than across lineages produces the results shown in Tables 7 and 8. Both for 10% contact and 20% contact, both the false alarm rate and detection rate decrease with mean retention rate, but increase with heterogeneity. Perhaps surprisingly, substantial heterogeneity in the retention rate does not significantly reduce the performance of the algorithm; indeed, the probability of false alarm is actually substantially reduced, with only moderate reduction in detection rate.

Table 6. Detection rate and false alarm rate as a function of retention rate — retention rate heterogeneous across lineages (80% nominal retention rate; 10% contact).

Retention Rate:	80% (nominal)	90% ± 5%	90% ± 10%	85% ± 5%	85% ± 10%	80% ± 5%	80% ± 10%
Detection Rate (%)	62.5	31.8	37.1	51.3	51.9	65.2	60.9
False Alarm Rate (%)	5.7	5.7	10.5	6.3	10.6	7.0	9.5

Table 7. Detection rate and false alarm rate as a function of retention rate — retention rate heterogeneous across characters (90% nominal retention rate; 10% contact).

Retention Rate:	90% (nominal)	90% ± 5%	90% ± 10%	85% ± 5%	85% ± 10%	80% ± 5%	80% ± 10%
Detection Rate (%)	42.4	44.7	45.9	52.3	50.5	59.3	54.1
False Alarm Rate (%)	6.0	4.2	2.6	4.7	3.1	5.4	3.0

Table 8. Detection rate and false alarm rate as a function of retention rate — retention rate heterogeneous across characters (80% nominal retention rate; 10% contact).

Retention Rate:	80% (nominal)	90% ± 5%	90% ± 10%	85% ± 5%	85% ± 10%	80% ± 5%	80% ± 10%
Detection Rate (%)	62.5	36.6	39.8	46.0	44.1	53.0	48.9
False Alarm Rate (%)	5.7	2.6	1.6	2.9	1.9	3.5	1.7

Experiment 4: In this experiment we consider the effects of multiple instances of contact on the performance. In addition to the case of a single instance of contact already considered, we distinguish five contact scenarios, each involving two instances of contact between languages in two sub-groups. The six contact scenarios are shown schematically in Figure 5(a). For example, Scenario #3 describes contact between a single donor language in one sub-group and two recipient languages in the second sub-group, while Scenario #4 describes contact between two donors in one sub-group and a single recipient in the other sub-group. In each instance of contact, we consider only unidirectional borrowing. For each scenario, pseudo-random data sets are generated using the parameter values specified previously in Table 4 except for the contact rate, which we set to 10% for one set of runs and then to 20% for a second set of runs.

The detection rate and false alarm rate are calculated using a threshold optimised for the Table 4 parameter values at 5% significance, summarized in Figure 5(b). The figure shows that performance varies depending on the contact scenario. At 10% contact, the detection rate varies between about 32% and 63%,

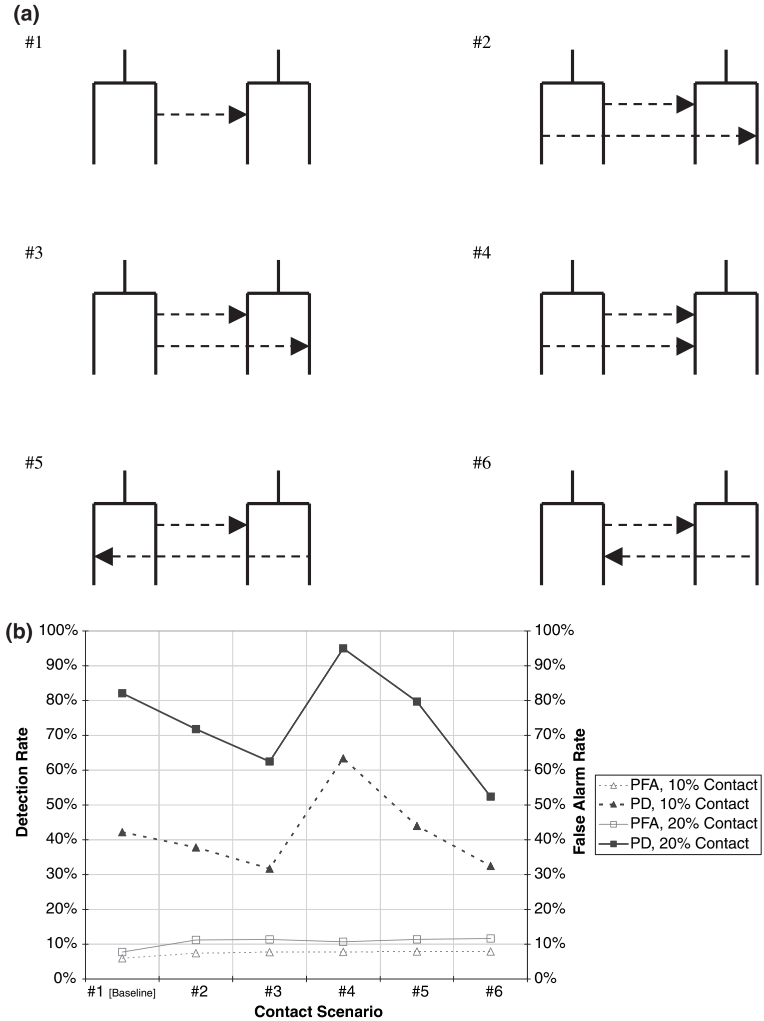


Figure 5(a). Six scenarios of contact between languages in two subgroups. Scenario #1 is the baseline scenario, for which there is only a single instance of contact. Scenarios #2 to #6 each involve two instances of contact.

Figure 5(b). Detection rate (PD) and false alarm rate (PFA) for the six language contact scenarios. (10% and 20% contact).

with a baseline rate of about 42%. The false alarm rate varies from the baseline rate of about 6% up to 8%. At 20% contact, the corresponding ranges are from 62% to 95% in the detection rate and 8% to 12% in the false alarm rate. This suggests that reasonable performance can be maintained with only slight increase in the false alarm rate in various contact situations.

It is evident that Scenario #4 offers the highest probability of correct detection for both 10% and 20% contact. In this scenario, the recipient language undergoes contact with two donor languages in the other sub-group. Each instance of contact causes the skewing between the recipient language and its siblings to increase, thereby substantially increasing the aggregate skewing of the recipient. The false alarm rate is increased slightly at both contact rates.

However, when the same donor language comes into contact with two recipient languages in the other sub-group, Scenario #3, the performance is substantially worse than that of the baseline scenario, Scenario #1. At both contact rates, the detection rate is decreased by more than 10% while the false alarm rate is slightly increased. The fall in the detection rate is caused by the two recipient languages having come into contact with the same donor, thereby reducing the skewing between them, and so also reducing the aggregate skewing. However, the skewing between other language pairs in the recipient sub-group is increased, causing the observed rise in the false alarm rate. The performance for Scenario #2 is similar, with the detection rate only slightly lower than that of the baseline scenario.

The greatest probability of false alarm is observed for Scenario #5. In this case, there is contact between a donor in each sub-group and a recipient in each sub-group. The aggregate skewing for the two recipient/donor pairs seem to cancel each other out to some degree, causing the detection rate to be approximately equal to that for the baseline scenario. For other language pairs, however, the bi-directional nature of the contact causes larger magnitude skewing values, both positive and negative, to occur more frequently. As a result the tails of the distribution of aggregate skewing are increased, causing the false alarm rate to increase. The performance for Scenario #6 is much the same, although the detection rate is somewhat lower.

The results of Experiment 4 show that, while two instances of contact between the two sub-groups often have a negative impact on performance, the test does still have practical value — indeed, when a single recipient language comes into contact with two donor languages, performance is substantially increased. We expect the test to have some utility for a greater number of instances of contact, although the performance will undoubtedly tend to decay.

4. Conclusion

The results of our experiments, which we discussed in the previous section, imply that Hinnebusch's method for inferring language contact based on lexical skewing can be useful in a number of situations. The formalization of his approach that we have presented here offers fairly robust performance at false alarm rates typically below 10%. In particular, 10% heterogeneity in the retention rate, both across characters and across lineages, does not substantially reduce the performance of the test. As Hinnebusch noted, while the method cannot be considered a replacement for a careful comparative study, it does provide an objective framework for generating plausible contact hypotheses to be probed in more depth.

A number of issues remain to be resolved, the main issue being its performance with actual data. In order to assess the practical level of performance, the method will have to be applied to several sets of languages for which the main instances of contact in each set are well-known — languages of the Indo-European family may provide at least one suitable test case.

The use of the performance measures derived here, based on synthetic data, to reflect the utility of the method in practice depend strongly on the realism of the language model adopted (see Appendix A in the supplementary material, available at http://www.philsoc.org.uk/transactions.asp). More realistic models of character replacement and contact, and models of multiple layers of contact-induced borrowing in particular, might also be examined. Nevertheless, the methods presented here provide a statistical study of the skewing method for detecting language contact introduced by Hinnebusch (1999). The method has been shown to be feasible for detecting language contact in a variety of simple scenarios and to be robust provided that the heterogeneity in the retention rate is not too great.

Dr James Minett
Department of Electronic Engineering
Ho Sin Hang Engineering Building
The Chinese University of Hong Kong
Shatin, N.T.
Hong Kong PRC
Email: jminett@ee.cuhk.edu.hk

Professor William S.-Y. Wang
Department of Electronic Engineering
Ho Sin Hang Engineering Building
The Chinese University of Hong Kong
Shatin, N.T.
Hong Kong PRC
Email: wsywang@ee.cuhk.edu.hk

REFERENCES

BANDELT, H.-J. & DRESS, A. W. M., 1992. 'A canonical decomposition theory for metrics on a finite set', *Adv. Math.* 92, 47–105.

BLUST, ROBERT, 2000. 'Why lexicostatistics doesn't work: the "universal" constant hypothesis and the Austronesian languages', in Colin Renfrew, April McMahon and Larry Trask (eds.), *Time Depth in Historical Linguistics*, Vol. 2, Cambridge: The McDonald Institute for Archaeological Research, 311—331.

BROOKS, DANIEL R., ERDEM, ESRA, MINETT, JAMES W. & RINGE, DONALD, 2005. 'Character-based cladistics and answer set programming', *Proceedings of the Seventh International Symposium on Practical Aspects of Declarative Languages (PADL 05)*, 37–51.

CAVALLI-SFORZA, LUIGI LUCA, MENOZZI, PAOLO & PIAZZA, ALBERTO, 1994. *The History and Geography of Human Genes*. Princeton: Princeton University Press.

CHEN, BAOYA, 2000. 'Relative rank analysis of core corresponding words in Sino-Tai', *Chinese Languages and Writings (Zhongguo Yuwen)* 277, 338–348.

DYEN, ISIDORE, KRUSKAL, JOSEPH B. & BLACK, PAUL, 1992. 'An Indoeuropean Classification: A Lexicostatistical Experiment', *Transactions of the American Philosophical Society* 82:5.

EMBLETON, SHEILA, 1981. *Incorporating Borrowing Rates in Lexicostatistical Tree Reconstruction*. Ph.D. dissertation, Department of Linguistics, University of Toronto.

Embleton, Sheila, 1986. *Statistics in Historical Linguistics*, Bochum: Brockmeyer.

Erdem, Esra, Lifschitz, V., Nakhleh, Luay & Ringe, Donald, 2003. 'Reconstructing the evolutionary history of Indo-European languages using answer set programming', *Proceedings of the Fifth International Symposium on Practical Aspects of Declarative Languages (PADL'03)*, 160–176.

Heine, Bernd, 1974. 'Historical linguistics and lexicostatistics in Africa', *Journal of African Linguistics* 11:3, 7–20.

Hinnebusch, Thomas J., 1999. '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.

Hinnebusch, Thomas J., 1999. 'Contact and lexicostatistics in comparative Bantu studies', in Jean-Marie Hombert and Larry M. Hyman (eds.), *Bantu Historical Linguistics: Theoretical and Empirical Perspectives*, Stanford, CA: CSLI Publications, 173–205.

Hübschmann, J. Heinrich, 1875. 'Über die Stellung des Armenischen im Kreise der indogermanischen Sprachen', *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete der Indogermanischen Sprachen*, 23:1, 5–49.

Minett, James W & Wang, William S-Y., 2003. 'On detecting borrowing: distance-based and character-based approaches', *Diachronica* 20:2, 289–330.

Ogura, Mieko & Wang, William S-Y., 1996. 'Lexical diffusion and evolution theory', in Raymond Hickey and Stanistaw Puppel (eds.), *Trends in Linguistics: Studies and Monographs 101. Language History and Linguistic Modelling: A Festschrift for Jacek Fisiak on his 60th Birthday*, Vol. 1. Berlin: Mouton de Gruyter, 1083–1098.

Renfrew, Colin, McMahon, April and Trask, Larry (eds.), 2000. *Time Depth in Historical Linguistics*, Vols. 1 & 2. Cambridge: The McDonald Institute for Archaeological Research.

Ringe, Don, Warnow, Tandy & Taylor, Ann, 2002. 'Indo-European and computational cladistics', *Transactions of the Philological Society* 100:1, 59–130.

Sagart, Laurent & Shixuan, Xu, 2001. 'History through loanwords: the loan correspondences between Hani and Chinese', *Cahiers de Linguistique — Asie Orientale* 30:1, 3–54.

Saitou, Naruya & Nei, Masatoshi, 1987. 'The neighbor-joining method: a new method for reconstructing phylogenetic trees', *Molecular and Biological Evolution* 4:4, 406–425.

Sankoff, David, 1972. 'Reconstructing the history and geography of an evolutionary tree', *American Mathematical Monthly* 79, 596–603.

Swadesh, Morris, 1951. 'Diffusional cumulation and archaic residue as historical explanations', *Southwestern Journal of Anthropology* 7, 339–346.

Swadesh, Morris, 1955. 'Towards greater accuracy in lexico-statistic dating', *International Journal of American linguistics* 18, 121–137.

Wang, Feng, 2004. *Language Contact and Language Comparison — the case of Bai*. Ph.D. thesis. City University of Hong Kong.

Wang, William S-Y, 1989. 'The migration of the Chinese people and the settlement of Taiwan', in Kwang-chih Chang, Kuang-chou Li, Arthur P. Wolf and Alexander Chien-chung Yin (eds.), *Anthropological Studies of the Taiwan Area: Accomplishments and Prospects*, Taipei: Department of Anthropology, National Taiwan University.

Wang, William S-Y, 1997. 'Languages or dialects?', *The CUHK Journal of Humanities* 1, 54–62.

Warnow, Tandy, Ringe, Donald & Taylor, Ann, 1995. 'Reconstructing the evolutionary history of natural languages', paper presented at the, *Workshop on Historical Linguistics*. University of Pennsylvania, 1995.

Wen, You, 1940. 'A study of synonyms in the Min-chia language', *Anthology of the Research Institute of Chinese Culture*, Huaxi Xiehe University, 1:1, 1–27 (in Chinese, Wen, You, 1940. ''Minjiayu zhong tongyizi zhi yanjiu''. *Zhongguo wenhua yanjiusuo jikan*, Huaxi xiehe daxue, 1:1, 1–27).

Xu, Tongqiang, 1991. *Historical Linguistics*, Beijing: The Commercial Press. (in Chinese, Xu, Tongqiang, 1991. *Lishi yuyanxue*, Beijing: Shangwu yinshuguan).