Detecting Borrowing Among Languages: a glottochronological approach

James W. Minett
William S.-Y. Wang
Language Engineering Laboratory
City University of Hong Kong

Introduction:

Linguistic classification allows the branching relationships of genetically related languages to be estimated, often represented by a tree. However, many of the methods used to reconstruct linguistic trees are only able to determine the vertical relationships that exist among a group of languages, i.e., the subgroupings of languages that are related by common descent from a proto-language. Horizontal relationships, which arise when languages come into contact and therefore tell us much about the history of languages and their speakers, can often not be detected by these methods and tend to distort the reconstructed linguistic trees (Dyen 1992).

A common method used to determine the vertical relationships among a group of languages is to determine lexical items that are *cognate* in each pair of languages; a *cognate* is a word that is related to a word in a sister language by common descent from the proto-language. Cognates can be difficult to distinguish from *borrowed words*, which are words that have been adopted by one language from other, possibly unrelated, languages. Cognates are indicative of vertical relationship while borrowed words are indicative of language contact and horizontal relationship.

Two approaches are often used to perform linguistic classification: the *comparative method*, which aims to establish sound correspondences that exist between two languages, and *lexicostatistics*, which works with the aggregate proportion of words that are cognate between two language. Both methods make use of lists of basic vocabulary (Swadesh 1951) that are supposed to be present in any language, regardless of culture, and are assumed to retain their form over significant time depths.

Problem Statement:

There are several criteria that can be used to distinguish cognates from borrowed words linguistically: phonological features, morphological complexity, the distribution of cognates, geographical & ecological clues, and semantics (Campbell 1998). Of these, the distribution of cognates is the most easily adopted for computational implementation.

Our aim is:

- to provide a method that can re-classify as borrowed words that had previously been classified as cognates;
- to develop a computational method for detecting borrowing so that more accurate reconstructions of the vertical relationships among groups of languages can be made.

We examine here two methods:

- The first method is an adaptation of lexicostatistics, in which languages are classified according to the lexical distance separating them — we attempt to detect borrowing by examining the reconstructed linguistic tree for branches with negative length, which we suggest indicates borrowing.
- The second method adopts the principle of parsimony and may be applied with the comparative method. Each possible linguistic tree is analyzed to determine the minimum number of changes of form required to explain the observed cognates, i.e. character congruence. Assuming that a particular form can originate only once, changes of form after the first are interpreted as borrowings.

A Concept for Detecting Borrowing:

In lexicostatistics, the lexical distance, LD_{ij}, between two languages can be defined in terms of the proportion of basic words that are cognate, C_{ij}, as

$$LD_{ij} = -\log(C_{ij}).$$

In the absence of borrowing,

- the proportion of the basic vocabulary retained by each language since splitting from the proto-language cannot exceed 100%;
- the lexical distance between the two languages cannot be negative.

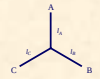
The sum of the lengths of the branches connecting two tips are constructed to model the lexical distance between the associated languages.

We therefore conclude that

Negative Branch Lengths ⇒ Borrowing

For any three languages for which the lexical distance, LD_{ij}, between them is known, a unique, unrooted binary tree can be constructed whose branch lengths are perfectly consistent with the lexical distances.

LD_{ij}	B	C
A	LD_{AB}	LD_{AC}
B	—	LD_{BC}



$$l_A = \tfrac{1}{2}(LD_{AB} + LD_{AC} - LD_{BC})$$

A sufficient and necessary condition for testing whether any branch length is negative is

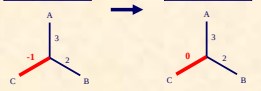
$$LD_{ij} + LD_{ik} < LD_{jk} \quad \text{for any distinct } i, j, k$$

An Algorithm for Detecting Borrowing:

The unrooted binary tree that represents the lexical distances given below-left has a branch with negative length connecting language **C** to the center. In the absence of borrowing, the branch length should be non-negative; the smallest change to the tree that can achieve this is to set the branch length to zero, as depicted below-right.

LD_{ij}	B	C
A	5	2
B	—	1

LD_{ij}	B	C
A	5	3
B	—	2



This suggests borrowing between language C and languages A or B

We propose the following algorithm algorithm:

1. Determine the lexical distance between all pairs of languages;
2. Examine all possible sub-groups of three languages:
   For each set of three languages, calculate its binary tree,
   Examine the tree for negative branch lengths;
3. If no tree has a negative branch length, EXIT,
   Otherwise, record the three languages, set the offending branch length to zero, and update the lexical distances,
   Goto 2.

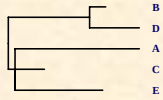
Example: Application to Synthetic Data

Observed Proportions of Cognates & Lexical Distance:

C_{ij} / LD_{ij}	A	B	C	D	E
A	—	60	80	50	80
B	5.10	—	90	85	75
C	2.23	1.05	—	90	75
D	6.93	1.63	1.05	—	70
E	2.23	2.88	2.88	3.56	—

Two taxa are implied by the data: (BCD) & (AE)

Reconstructed Tree:



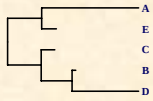
The expected taxa are not reflected in the optimal topology

Proportions of Cognates & Lexical Distance after running borrowing detection algorithm:

C_{ij} / LD_{ij}	A	B	C	D	E
A	—	59.8	66.7	50	75.6
B	5.16	—	89.1	83.7	75
C	4.05	1.15	—	75.0	75
D	6.93	1.77	2.88	—	66.1
E	2.80	2.88	2.88	4.13	—

Two pairs of languages, (A,C) and (C,D), are set to be lexically more distant

Reconstructed Tree:



Both expected taxa, (BCD) & (AE), are recovered

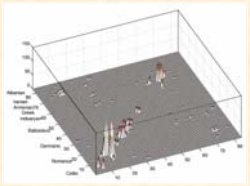
Test Case — Indoeuropean:

Dyen, Kruskal & Black (1992) published a lexicostatistical classification of the Indoeuropean family for 84 word lists, each list representing one dialect. They state that "a … special problem that can arise is the inflation of percentages due to borrowing that has not been specifically detected". We have run our algorithm on their entire data set to find out whether we can detect instances of undetected borrowing between lists.

The figure below shows graphically the difference between the input cognate matrix and the updated cognate matrix after application of the algorithm. Values of this matrix indicate the proportion of words borrowed between each pair of lists. More than 50 borrowing events were indicated, including :

Dominican French Creole [16] & French [13] (12.3%); Nepali [64] & Khaskura [65] (6.4%); Afrikaans [27] & Dutch [26] (4.4%)

Borrowing matrix for the Indoeuropean data set:



Two types of feature are immediately apparent:

- borrowing is often indicated among closely related languages, particularly evident in Romance, Germanic, & Indoaryan.
- borrowing also appears to be indicated between subfamilies, most evident between French dialects [12-16] and Iranian [73-75] — we examine this case in more detail

The case of French and Iranian:

The borrowing matrix indicates borrowing among the following lists:
French, Provençal, Walloon, French Creole (2), Ossetic, Waziri, and Afghan.
For this set of lists only, the borrowing events indicated by the algorithm are:

French	↔	French Creole C	**OR**	Waziri
Provençal	↔	French Creole C	**OR**	Waziri
Walloon	↔	French Creole C	**OR**	Waziri
French	↔	French Creole D	**OR**	Ossetic
Walloon	↔	French Creole D	**OR**	Ossetic
Provençal	↔	French Creole D	**OR**	Ossetic
French	↔	French Creole C	**OR**	Afghan
Provençal	↔	French Creole D	**OR**	Afghan
Walloon	↔	French Creole C	**OR**	Afghan
French	↔	French Creole D	**OR**	Afghan
Provençal	↔	French Creole C	**OR**	Afghan
French Creole D	↔	French	**OR**	French Creole C

In each case, borrowing is indicated between a French Creole and either a French dialect or an Iranian language, borrowing from the French dialects into the French Creoles being by far the most likely explanation. This result reflects the status of French Creole to its superstrate language, French.

Borrowing is indicated between French Creole and the French dialects

The presence of negative branch lengths in the trees constructed for the triplets listed above is the only indication of borrowing between the French Creoles and dialects — the algorithm is unable to detect evidence for borrowing among them for any other triplets. The linguistic reason for borrowing being evident when comparing French with Iranian is so far unclear.

An Alternative Proposal for Detecting Borrowing:

Hennig (1950) developed the method of *phylogenetic systematics* to perform hierarchical classification of organisms. In his method, organisms are classified according to the shared states of various characters; only shared innovations (synapomorphies) are indicative of common ancestry (Scotland 1992). The subgrouping that requires the fewest number of innovations is selected as the most parsimonious classification. The same method was developed independently by Platnick and Cameron (1977) for application to linguistic classification. The method can be used effectively in conjunction with the comparative method.

We have adopted the same principle of parsimony to distinguish horizontally transmitted linguistic characters from vertically transmitted linguistic characters; the features can be lexical, syntactical or morphological — we have worked with lexical features, in particular, with Swadesh lists of basic meanings.

To demonstrate, consider two topologies, shown on the right, for a group of 7 hypothetical languages having a particular character with either of two states.

Which topology is more parsimonious?

Topology **A** requires 2 state changes for congruence, whether state **1** or state **2** is the retained state. Topology **B** is equally parsimonious for retained state **1**, but is less parsimonious for retained state **2**, requiring 3 state changes.

In our experiments, we have assumed that the retained state is that which is observed in the most taxa; thus for topology **A**, state **1** is observed in 2 taxa while state **2** is observed in 3 taxa; state **2** is therefore assumed to be the retained state.

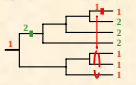
For a set of characters, the optimal topology is that requiring the fewest total number of state changes.

How to detect borrowings?

We assume that a particular innovation can occur only once; all further changes to that state are assumed to be due to borrowing.

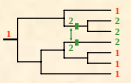
We therefore examine the optimal topologies for states for which more than one state change is required — **all but the first state change are borrowings.**

Topology A – Retained State 1:



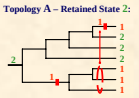
1 innovation, 1 borrowing

Topology B – Retained State 1:



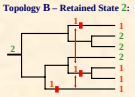
1 innovation, 1 borrowing

Topology A – Retained State 2:



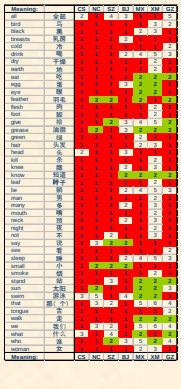
1 innovation, 1 borrowing

Topology B – Retained State 2:



1 innovation, 2 borrowings

Application to the Seven Main Dialects of Chinese — 3 candidate topologies, 1 random topology:



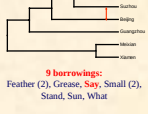
Topology 1:

7 borrowings:
Feather (2), Grease, **Say**, Small, Sun, What

Topology 2:

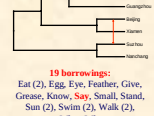
8 borrowings:
Feather (2), Grease, Know, **Say**, Small, Sun, What

Topology 3:



9 borrowings:
Feather (2), Grease, **Say**, Small (2), Stand, Sun, What

Topology 4 (random):



19 borrowings:
Eat (2), Egg, Eye, Feather, Give, Grease, Know, **Say**, Small, Stand, Sun (2), Swim (2), Walk (2), What, Who

Conclusion: **Topology 1 (with 7 borrowing events) is the most parsimonious classification. The method is able to distinguish horizontal from vertical transmission.**

Conclusion:

We have demonstrated two proposed methods for distinguishing horizontally transmitted linguistic feature from vertically transmitted features: the first based on detecting branches with negative length in trees constructed using lexicostatistics, the second based on determining the most parsimonious topology in terms of the number state changes to achieve congruence.

The first method has been applied to detect borrowing among 84 dialects of Indoeuropean, indicating borrowing within several subfamilies and, in particular, between French Creole and the French dialects. However, its inability to distinguish the words that have been borrowed is a serious weakness.

The second method, which has been applied in a preliminary way to the seven main dialects of Chinese, has been far more successful. The three topologies considered candidates for the best linguistic tree are shown to be similarly parsimonious; of these, the topology containing the taxa (Changsha, Nanchang, Suzhou, Beijing) and (Guangzhou, Meixian, Xiamen) is shown to be optimal. Moreover, the borrowed words are identified, for the most part consistently. The fourth topology, chosen at random, is shown to be far less parsimonious and should be rejected as a potential classification of the seven dialects.

We therefore believe the second method based on parsimony should be investigated in more detail.

References:

Bergsland, Knut and Hans Vogt (1962) "On the validity of glottochronology," *Current Anthropology* 3: 115–53.

Campbell, Lyle (1998) Historical Linguistics: An Introduction, Edinburgh University Press, Edinburgh.

Dyen, Isodore, Joseph B. Kruskal, and Paul Black (1992) "An Indoeuropean classification: A lexicostatistic experiment," *Transactions of the American Philosophical Society*, Volume 82, Part 5, 1992.

Embleton, S. M. (1982) "Lexicostatistical tree reconstruction incorporating borrowing," Eighth LACUS Forum, Columbia: Hornbeam, pp. 265-272.

Hennig, Willi (1950) Grundzüge einer Theorie der phylogenetischen Systematik, Deutsche Zentralverlag, Berlin.

O'Hara, Robert J. (1996) "Trees of history in systematics and philology," *Memorie della Società Italiana di Scienze Naturali e del Museo Civico di Storia Naturale di Milano, 27(1):* 81-88.

Platnick, N. I. and H. D. Cameron (1977) "Cladistic methods in textual, linguistic, and phylogenetic analysis," *Systematic Zoology* 26: 380–5.

Scotland, R. W. (1992) "Cladistic theory," in Forey, Peter L. et al. Cladistics: A Practical Course in Systematics, Clarendon Press, Oxford.

Swadesh, Morris (1951) "Diffusional cumulation and archaic residue as historical explanations," *Southwest Journal of Anthropology*, 7: 1–21.

Evolution of Language, Harvard University, March 27–31, 2002