

# On the Detection of Borrowing Among Languages



James W. Minett  
*City University of Hong Kong*

November 23, 2001



# Introduction

- ***Borrowing*** is the process by which “speakers of one language adopt elements of another” <sup>1</sup>
- ***Problem:***  
Undetected borrowing can cause languages to appear to be more closely related genetically than is the case
- ***Aim:***  
To develop a method for detecting borrowing so that more accurate reconstructions of hierarchical relationships among languages can be made.

<sup>1</sup> (Lehmann 1992b)



# Lexicostatistics

- **Lexicostatistics** is a method for reconstructing the hierarchical relationships among a group of genetically related languages <sup>1</sup>.
- **Basic vocabulary** of each language is modeled by a word list whose meanings are considered culturally universal <sup>1</sup>
  - e.g. head, person, fish, sun
- “Two forms are **cognate** if they have both descended in unbroken lines from the same ancestor” <sup>2</sup>
  - e.g. English “*night*” and German “*Nacht*” are cognates,  
English “*head*” and German “*Kopf*” are not <sup>3</sup>.

<sup>1</sup> (Embleton 1986)   <sup>2</sup> (Dyen, Kruskal & Black 1992)   <sup>3</sup> (Lehmann 1992b)



# Lexicostatistics

- The lexicostatistical method applied to a group of languages <sup>1</sup>:
  - collect a word list for each language;
  - determine cognate words shared by each pair of languages;
  - calculate the lexicostatistical percentages, i.e. the proportion of cognates, for each pair of languages;
  - subgroup the languages in some way based on the lexicostatistical percentages,
    - pair-group <sup>2</sup>, neighbor-joining <sup>3</sup>, percent standard deviation <sup>4</sup>, exhaustive search <sup>5</sup>

<sup>1</sup> (Dyen, Kruskal & Black 1992)   <sup>2</sup> (Sneath & Sokal 1973)   <sup>3</sup> (Saitou & Nei 1992)

<sup>4</sup> (Fitch & Margoliash 1992)   <sup>5</sup> (Qiao & Wang 1992)

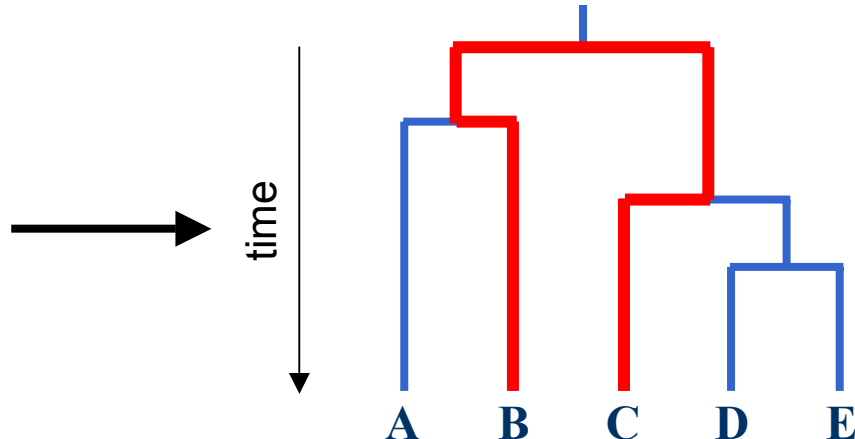
# Lexicostatistical Subgrouping

- Define a **lexical distance** between each pair of languages

$$LD_{i,j} = -\log(C_{i,j})$$

- Aim is to construct a rooted binary tree that “best” fits the lexical distances among a **group** of languages:

$LD_{ij}$	A	B	C	D	E
A	0	x	x	x	x
B	x	0	x	x	x
C	x	x	0	x	x
D	x	x	x	0	x
E	x	x	x	x	0





# Borrowing & Lexicostatistics

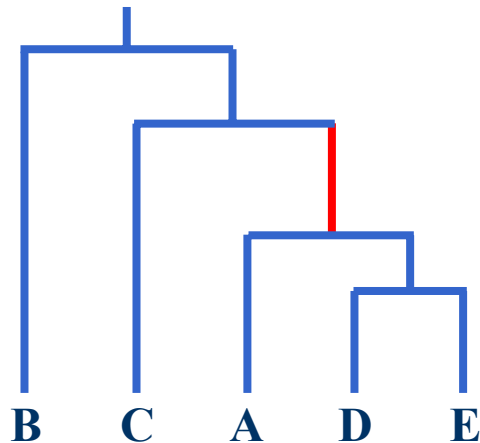
- ***Borrowing*** is the process by which “speakers of one language adopt elements of another”<sup>1</sup>
  - e.g. English “*flower*” borrowed from French “*fleur*”<sup>2</sup>
- Undetected borrowing can cause inflation of lexicostatistical percentages<sup>2</sup>:
  - languages may appear to be more closely related genetically than is the case
  - lexicostatistical subgrouping of the languages may be affected

<sup>1</sup> (Lehmann 1992b)    <sup>2</sup> (Dyen, Kruskal & Black 1992)

# Undetected Borrowing

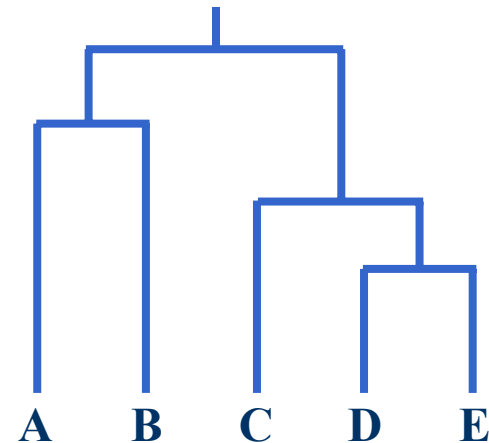
Borrowing undetected:

$c_{ij}$	B	C	D	E
A	72	<b>89</b>	58	66
B	-	62	52	52
C	-	-	62	65
D	-	-	-	72



Borrowing detected:

$c_{ij}$	B	C	D	E
A	72	<b>69</b>	58	66
B	-	62	52	52
C	-	-	62	65
D	-	-	-	72

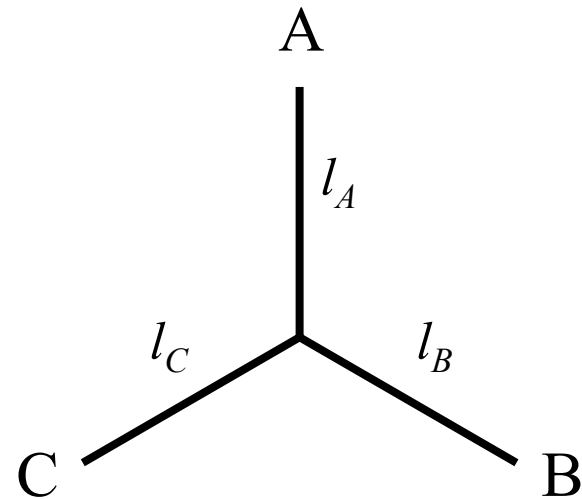


# Constructing a Binary Tree for 3 Languages

given the matrix of  
lexical distances:

$LD_{ij}$	B	C
A	$LD_{A,B}$	$LD_{A,C}$
B	—	$LD_{B,C}$

calculate the unique  
unrooted binary tree:



$$l_A = \frac{1}{2} (LD_{A,B} + LD_{A,C} - LD_{B,C})$$



# Negative Branch Lengths

Words in the basic vocabulary of a language are lost over time

In the absence of borrowing:

- The proportion of the basic vocabulary of a language retained along any branch of the tree  $\leq 100\%$
- Lexical distance along any branch cannot be negative

Implication:

- Negative branch lengths are due to borrowing



# A Condition for Detecting Borrowing

- A negative branch length occurs when

$$l_i = \frac{1}{2} (LD_{i,j} + LD_{i,k} - LD_{j,k}) < 0 \quad \text{for any distinct } i, j, k$$

- An equivalent condition is

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

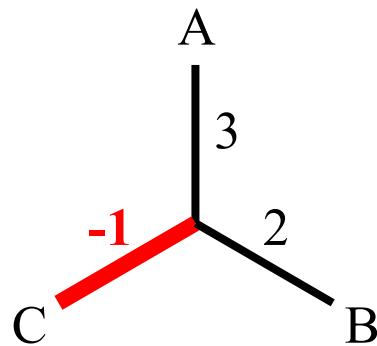
or, in terms of the lexicostatistical percentages,

$$C_{i,j} + C_{i,k} > C_{j,k} \quad \text{for any distinct } i, j, k$$

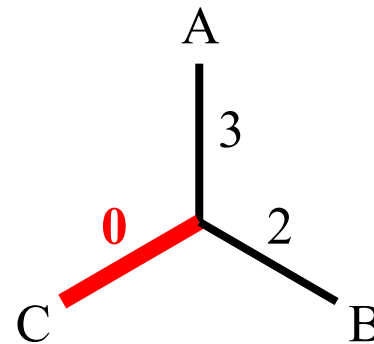
# Which language did the borrowing?

- In order to recover a “valid” tree with all branch lengths  $\geq 0$ , set the length of the branch with negative length to zero

$LD_{ij}$	B	C
A	5	2
B	—	1



$LD_{ij}$	B	C
A	5	<b>3</b>
B	—	<b>2</b>



- Suggests borrowing between C and either A or B, or both



## Why Bother?

- ***Aim:***  
To develop a method for detecting borrowing so that more accurate reconstructions of hierarchical relationships among languages can be made.
- May not have sufficient data to distinguish loan words from cognates
- May have incorrectly classified some loan words as cognates
- An algorithm that can produce reasonably accurate subgroupings without detailed treatment of borrowing would be helpful



# An Algorithm for “Correcting” Borrowing

1. Determine the lexical distances between all pairs of languages.
2. Examine all groups of three languages for negative branch lengths.
3. If at least one branch length is negative:
  - set the largest magnitude negative branch length to zero,  
return to step 1.

Otherwise, exit.

# An Example

$LD_{ij} \backslash C_{ij}$	A	B	C	D	E
A	—	60	80	50	80
B	5.1	—	90	85	75
C	2.2	1.1	—	90	75
D	6.9	1.6	1.1	—	70
E	2.2	2.9	2.9	3.6	—

## An Example: one of the sub-trees

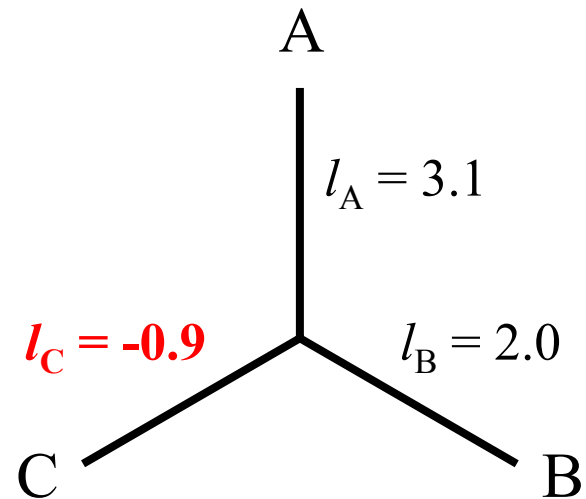
- Consider the tree for group (A,B,C):

$LD_{ij}$	B	C
A	<b>5.1</b>	<b>2.2</b>
B	—	<b>1.1</b>

Note that

$$LD_{B,C} + LD_{A,C} < LD_{A,B}$$

suggests borrowing  $A, B \leftrightarrow C$



Also note the negative branch length,  $l_C$

## An Example: the first pass of the algorithm

- The following trees have a negative branch length:

$$(A,C,D) \quad l_C = -1.8 \quad \Rightarrow \quad A \text{ and/or } D \leftrightarrow C$$

$$(A,B,C) \quad l_C = -0.9 \quad \Rightarrow \quad A \text{ and/or } B \leftrightarrow C$$

$$(A,D,E) \quad l_E = -0.6 \quad \Rightarrow \quad A \text{ and/or } D \leftrightarrow E$$

$$(A,B,D) \quad l_B = -0.1 \quad \Rightarrow \quad A \text{ and/or } D \leftrightarrow B$$

- The largest magnitude negative branch length is

$$l_C = -1.8 \text{ in tree } (A,C,D)$$

so we set that branch length to zero and recalculate the distance matrix

## An Example: “corrected” distance matrix

- The algorithm makes the following corrections:
  1.  $A, D \leftrightarrow C$
  2.  $A, D \leftrightarrow E$
  3.  $C, D \leftrightarrow B$
  4.  $A, D \leftrightarrow B$
- Distance matrix is changed (significant changes in red):

$LD_{ij}$	A	B	C	D
B	5.1	—	—	—
C	2.2	1.1	—	—
D	6.9	1.6	1.1	—
E	2.2	2.9	2.9	3.6

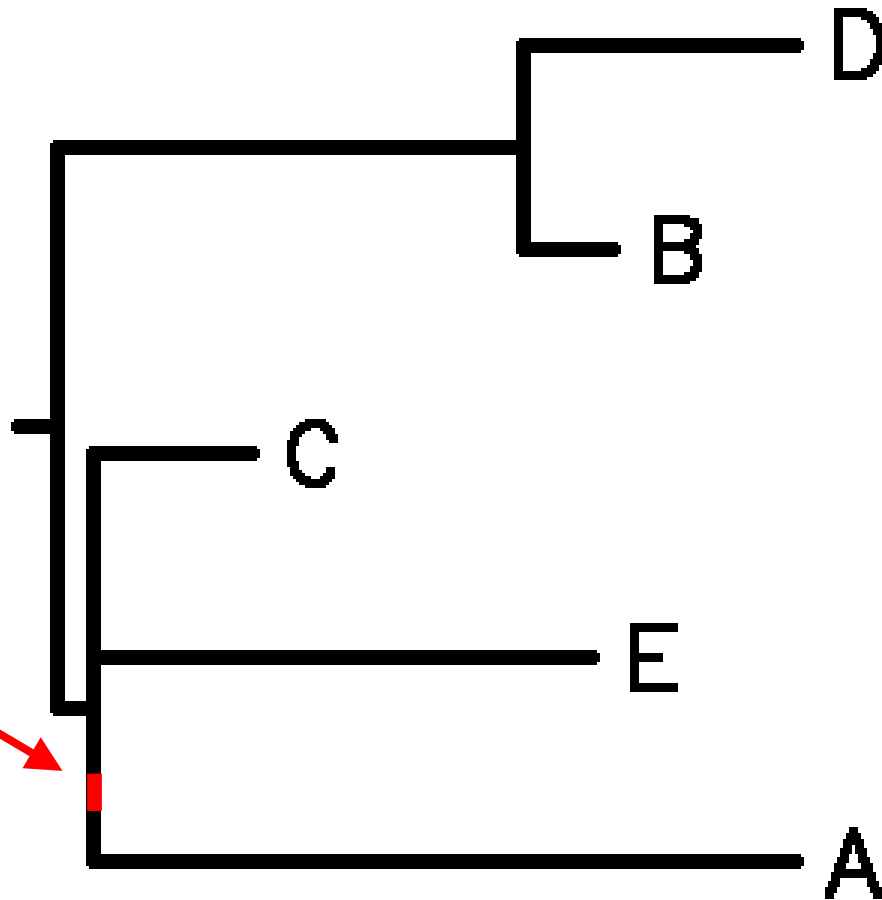


$LD_{ij}$	A	B	C	D
B	5.2	—	—	—
C	4.1	1.2	—	—
D	6.9	1.8	2.9	—
E	2.8	2.9	2.9	4.1

# An Example: optimal tree before “correction”

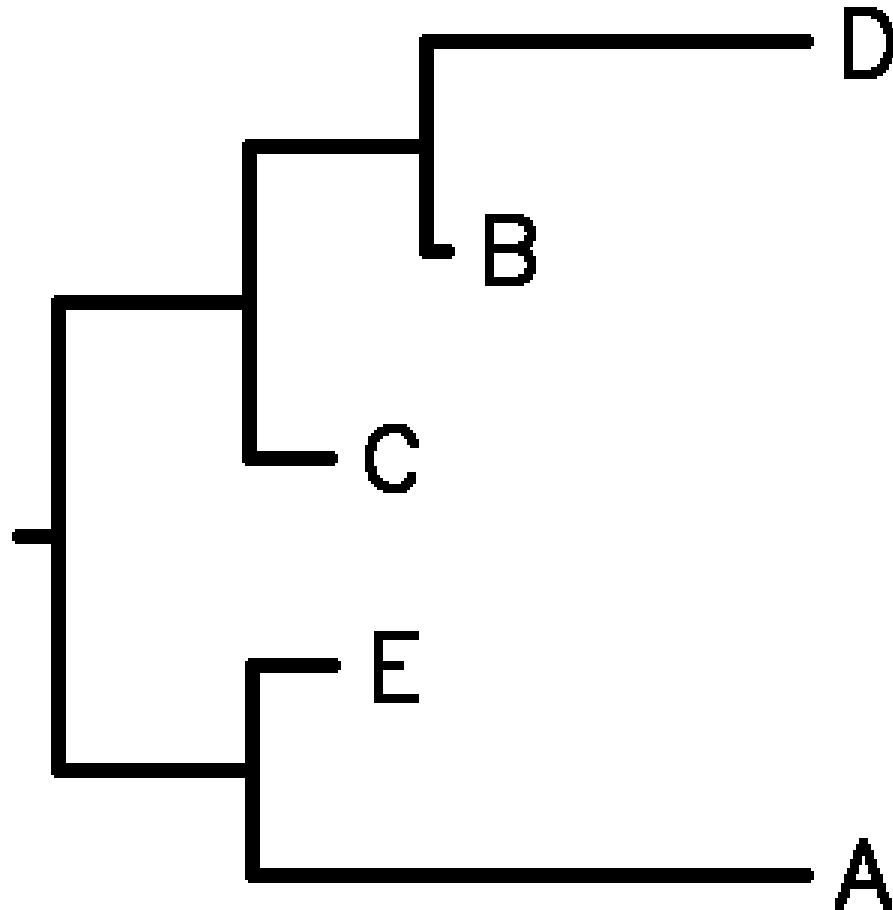
The languages B, C, & D  
are not shown as a  
monophyletic group

negative branch length



# An Example: optimal tree after “correction”

The languages B, C, & D  
are shown as a  
monophyletic group



Optimal tree obtained using a least-mean-squared exhaustive search algorithm (Qiao & Wang 1998)



## Application to Indo-European Languages

- Applied algorithm to a collection of 84 word lists of Indo-European languages collated by Dyen et al. <sup>1</sup>
- Dyen writes “A ... special problem that can arise is the inflation of percentages due to borrowing that has not been specifically detected,” and continues “Errors of this kind are only likely to occur between closely related dialects,” citing Nepali and Hindi as examples.
- Borrowing indicated between 50+ word list pairs, including:
  - Dominican French Creole & French 12.3%
  - Nepali & Khaskura 6.4%
  - Afrikaans & Dutch 4.4%
  - Spanish & Portuguese 2.0%
  - Icelandic & Faroese 2.0%

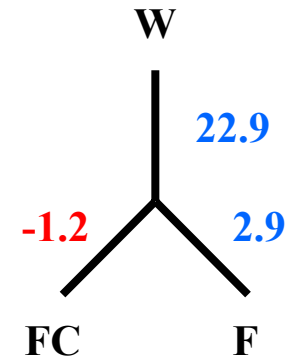
<sup>1</sup> (Dyen, 1992)

# The First Pass of the Algorithm

- Borrowing indicated among:  
French Creole (Dominican), French, Waziri

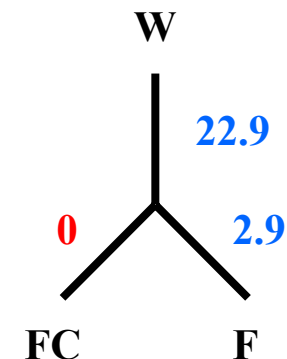
BEFORE:

	FC	F	W
FC	—	84.7	11.4
F	1.7	—	7.6
W	21.7	25.8	—

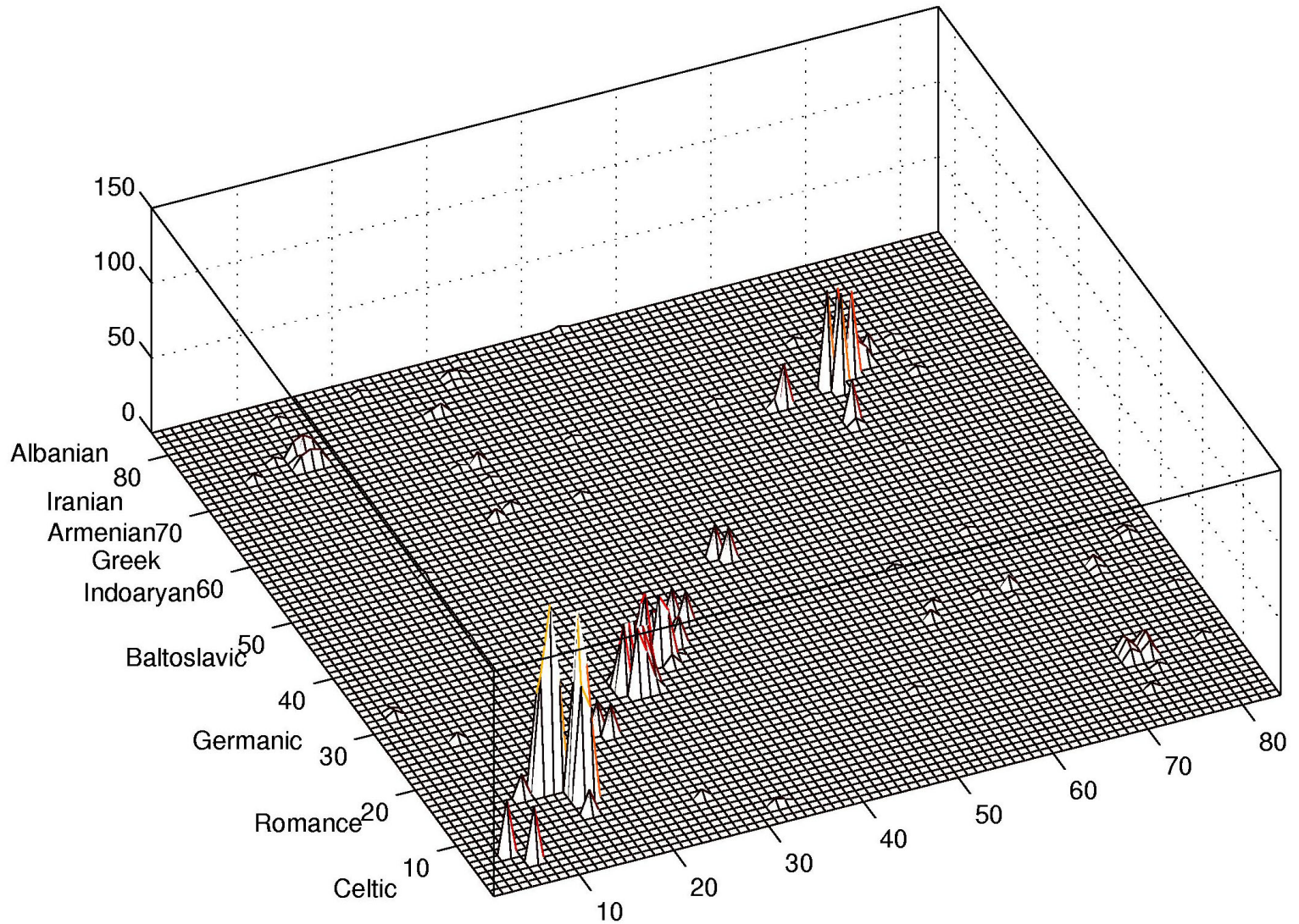


AFTER:

	FC	F	W
FC	—	75.2	10.1
F	2.9	—	7.6
W	22.9	25.8	—



# Borrowing Indicated Among IE Languages





## Future Work

- Correction algorithm is ad hoc:—  
should develop formal basis for the algorithm, or a new algorithm
- Borrowing indicated by negative branch lengths only:—  
should improve resolution
- Only able to detect borrowing in large data sets:—  
should test more data
- Should relate to linguistic knowledge of language groups  
to confirm practical validity of algorithm
- Should relate to other lexicostatistical methods, e.g. Embleton<sup>1</sup>

<sup>1</sup> (Embleton1982)



## Summary

- Use negative branch lengths as a means to detect borrowing
- Developed algorithm to detect and “correct” borrowing among a group of languages
- For the Indo-European languages, borrowing apparently detected between 50+ pairs of dialects
- Some borrowing appears to be detectable for large groups but not for small groups



# References

- Dyen I., Kruskal J. B., and Black P. 1992. "An Indo-European classification: a lexicostatistical experiment," *Transactions of the American Philosophical Society*, Volume 82, Part 5.
- Embleton, S. M. 1982. "Lexicostatistical tree reconstruction incorporating borrowing," *Eighth LACUS Forum*, Columbia: Hornbeam, pp. 265-272.
- Embleton, S. M. 1986. "Statistics in historical linguistics," *Quantitative Linguistics*, Vol. 30, Bochum: Studienverlag Brockmeyer.
- Felsenstein, J. 1980. *Phylip: Phylogeny Inference Package*.
- Fitch W. and Margoliash E. 1967. "Construction of phylogenetic trees," *Science*, Vol. 155, pp. 279-284.
- Krishnamurti, Bh., Moses, L., and Danforth, D. G. 1983. "Unchanged cognates as a criterion in linguistic subgrouping," *Language*, Vol. 39, No. 3, pp. 541-568.
- Lehmann, W. P. 1992a. *Historical linguistics: an introduction*, 3rd ed., London: Routledge.
- Lehmann, W. P. 1992b. *Workbook for Historical Linguistics*, 3rd ed., Dallas: Summer Institute of Linguistics, pp. 67-73.
- Qiao, S. and Wang W. S.-Y. 1998. "Evaluating phylogenetic trees by matrix decomposition," *Anthropological Science*, Vol. 106, No. 1, pp. 1-22.
- Saitou N. and Nei. M. 1987. "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Molecular Biology and Evolution*, Vol. 4, pp. 406-425.
- Sneath P. A. and Sokal R. R. 1973. *Numerical Taxonomy*, San Francisco: W. H. Freeman.
- Swadesh, M. 1951. "Diffusional cumulation and archaic residue as historical explanations," *Southwest Journal of Anthropology*, Vol. 7, pp. 1-21.