

# **Vertical or Horizontal?**

## **Distinguishing Modes of Transmission**

**James W. Minett**  
**Wang Feng**

**Language Engineering Laboratory**  
**City University of Hong Kong**

# Outline

- **Key stages in linguistic classification**
- **Modes of transmission — vertical and horizontal**
- **An cladistic method for distinguishing vertical transmission from horizontal transmission**
- **Application to Chinese dialects**
- **Conclusion**

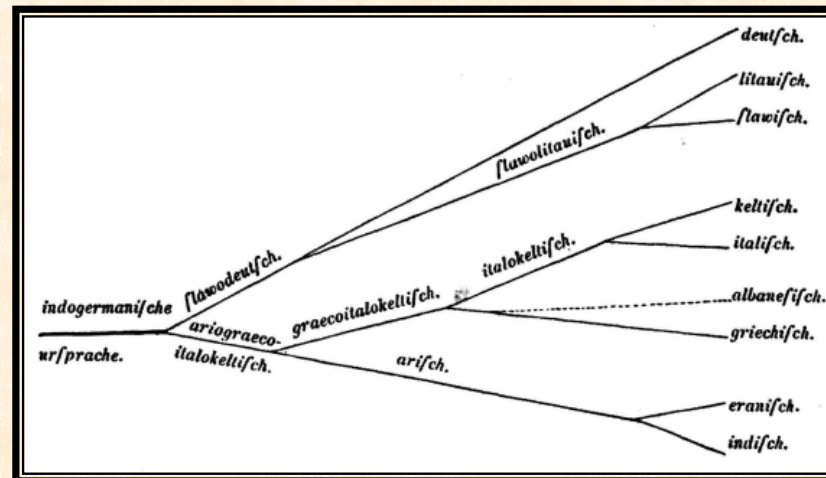
# Key Stages in Linguistic Classification

- **William Jones, 1786:**
  - Jones noted that Sanskrit, Greek and Latin “have sprung from some common source which, perhaps, no longer exists”
  - this paved the way for hierarchical classification of groups of languages
- **August Schleicher, 1861:**
  - the first to use family tree diagrams to represent the hierarchical relationships among languages
- **Hennig, 1950:**
  - aimed to devise a method for discovering the ancestor-descendant relationships among biological species implied by Darwin’s theory of evolution
  - the same method, *cladistics*, can be applied to linguistic classification

# Vertical Transmission

- **Vertical Transmission:** the transmission of a linguistic feature from one *generation* of speakers to the next
- Independent changes in the linguistic features of different groups of speakers gives rise to *increasingly* distinct dialects & languages
- In classification, two languages should be grouped together if they are *more closely related* to each other than to any other language

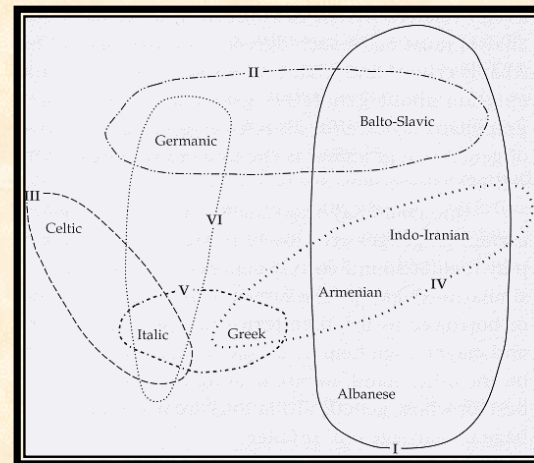
Schleicher's tree for Indo-European (1861)



# Horizontal Transmission (Borrowing)

- **Horizontal Transmission: the transmission of a linguistic feature from one *group* of speakers to another**
- **Diffusion of linguistic features among languages that come into contact gives rise to *decreasingly* distinct languages**
- **Failing to distinguish *retained* features (vertical) from *borrowed* features (horizontal) will lead to poor classifications**

Some overlapping linguistic features in a linguistic area – Indo-European (Bloomfield 1933)



# The Comparative Method (Ross 1996)

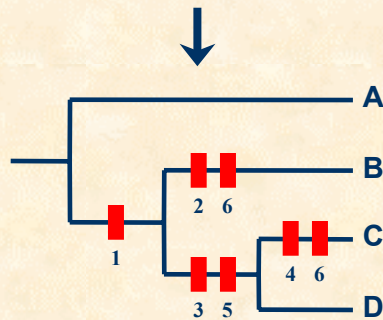
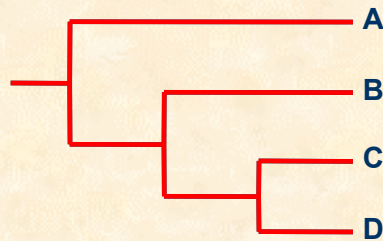
- **Determine that the set of languages are genetically related**
- **Collect putative cognates sets for each language**
  - a pair of words in two languages are cognate if they share (essentially) the same meaning and derive from a common form (by vertical transmission)
- **Determine sound correspondences among cognates sets**
  - e.g. Grimm's Law describes sound changes in the transition from Proto-Indo-European to Proto-Germanic:  $b, d, g \rightarrow p, t, k$  etc.
- **Reconstruct protolanguage of the family**
- **Determine innovations shared by sub-groups of languages**
- **Construct the “family tree”**

... but what if we fail to detect borrowing?

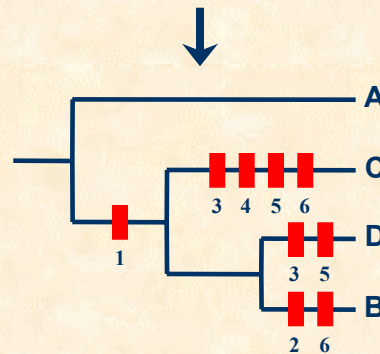
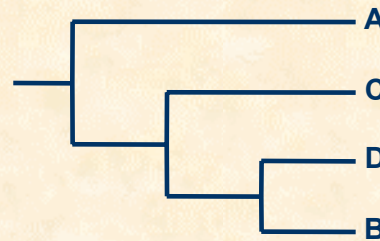
# The Cladistic Method (Kitching 1998)

	Characters					
Taxa	1	2	3	4	5	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

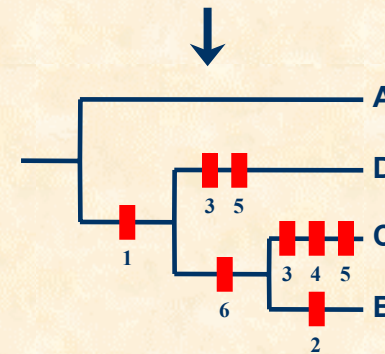
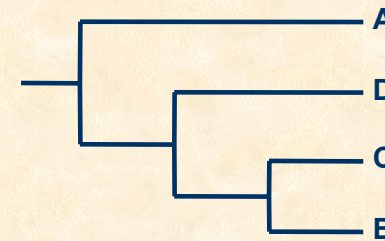
Example from  
(Kitching et al. 1998)



7 state changes



9 state changes



8 state changes

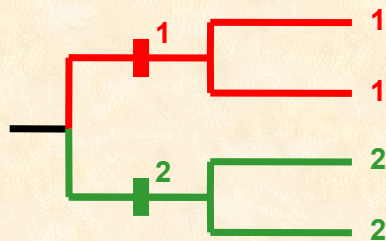
# Detecting Borrowing

- **Assumption:**

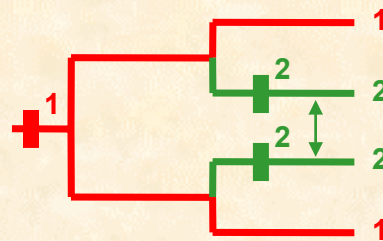
- no character state emerges independently more than once
- this is *probably not valid* for sound correspondences; identical sound changes may occur independently in unrelated languages, e.g. voiced stop finals changing to unvoiced stop finals
- but it *may be valid* for lexical items (meaning–form pairs); similar forms for a “basic” meaning are likely to be cognates or borrowings

- **Implication:**

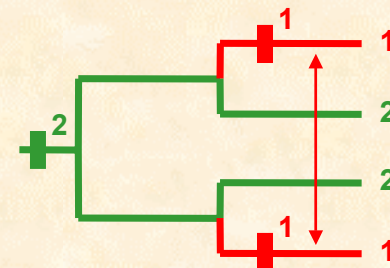
- second and further instances of a change of character state suggest borrowing



no borrowing



borrowing



borrowing

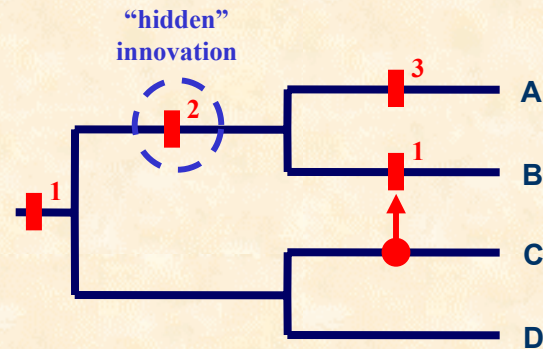
# Detecting Borrowing

- **This suggests the following simple algorithm for performing genetic classification while simultaneously detecting borrowing:**
  - perform cladistic analysis to determine most parsimonious trees
  - for each tree topology and each character, count the total number of state changes for each meaning
  - each state change beyond the first is recorded as one borrowing
  - select the most parsimonious trees
  - for each selected tree and each character, determine the most likely borrowings

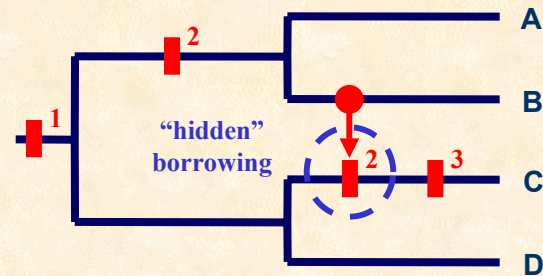
(record all consistent combinations of borrowing)

# Some Features are Hidden

- We cannot detect “hidden” innovations:



- We cannot detect “hidden” borrowings:

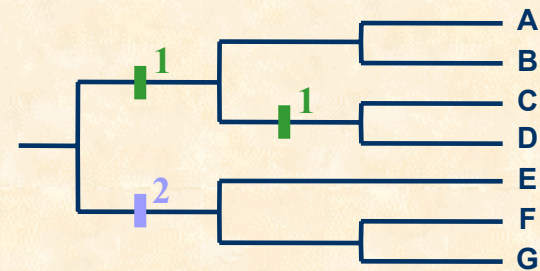


# A Simple Test of the Algorithm

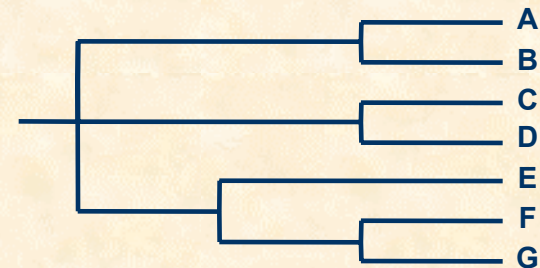
- A synthetic character set with consistent characters:

Characters	Languages						
	A	B	C	D	E	F	G
1	0	0	0	0	0	0	0
2	0	0	0	0	1	1	1
3	1	1	1	1	0	0	0
4	0	0	1	1	2	2	2
5	1	1	0	0	2	2	2
6	1	1	1	1	0	2	2
7	1	1	1	1	2	0	0
8	0	0	1	1	2	3	3
9	1	1	0	0	2	3	3
10	1	1	2	2	0	3	3
11	1	1	2	2	3	0	0
12	0	1	2	2	3	4	4
13	1	1	0	2	3	4	4
14	1	1	2	2	0	3	4
15	0	1	2	3	4	5	5
16	1	2	0	3	4	5	6
17	1	2	3	3	0	4	5
18	1	0	2	2	3	3	3
19	1	2	3	0	4	5	6
20	1	2	3	4	5	6	7

Actual topology:



Reconstructed consensus tree:



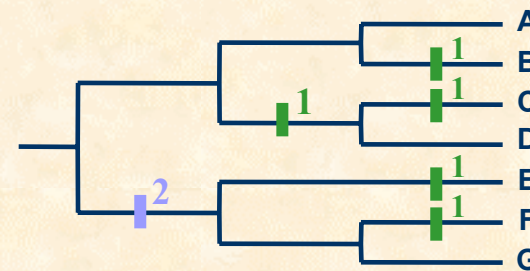
41 trees, including the actual topology, are congruent with the data

# A More Complex Test of the Algorithm

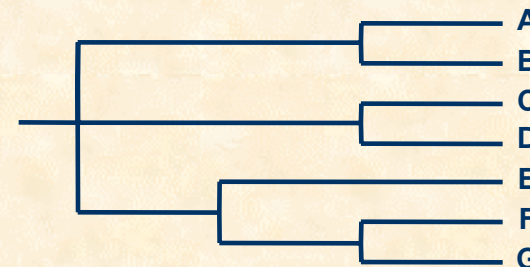
- A synthetic character set with one conflicting character:

Characters	Languages						
	A	B	C	D	E	F	G
1	0	0	0	0	1	1	1
2	1	1	1	1	0	0	0
3	0	0	1	1	2	2	2
4	1	1	0	0	2	2	2
5	1	1	1	1	0	2	2
6	1	1	1	1	2	0	0
7	0	0	1	1	2	3	3
8	1	1	0	0	2	3	3
9	1	1	2	2	0	3	3
10	1	1	2	2	3	0	0
11	1	0	0	2	3	3	3
12	1	0	0	2	0	3	3
13	1	0	2	3	0	0	4
14	2	1	3	4	0	0	5
15	0	0	1	2	0	3	4
16	1	2	0	0	3	4	4
17	1	0	3	4	0	5	5
18	1	1	0	0	0	0	0
19	0	0	1	1	0	0	0
20	0	1	1	0	1	1	0

Actual topology:



Reconstructed consensus tree:



7 trees, including the actual topology,  
each indicate 2 borrowings  
(C20 — 2 borrowings among B, C, E, & F)

# Questions

- **How to select the “best” genetic tree among the most parsimonious trees?**
  - could construct a consensus tree
  - instead, examine each of the most parsimonious trees for repeated structures
  - use simple hypothesis testing to determine hierarchy

structures that have a non-negligible probability of being observed by chance in a random sample are rejected (c.f. Baxter (2000))

# Questions

- **How to identify the donor and receiver languages?**
  - for each of the most parsimonious trees and for each borrowed meaning, list all possible combinations of donor and receiver language
  - **the linguistic must make the final judgment**

# Application to Chinese

- **Seven main dialects of Chinese:**

Mandarin (Beijing), Xiang (Changsha),  
Yue (Cantonese), Gan (Nanchang),  
Wu (Suzhou), Hakka (Meixian), Min (Xiamen)

- **How should they be genetically classified?**

there are 10,395 potential distinct binary tree classifications

- **Can we identify which meanings have been borrowed?**

for the Swadesh 100 list of meanings (Swadesh 1951),  
only 45 meanings are diagnostic for the Chinese dialects  
(data adapted from Xu (1991))

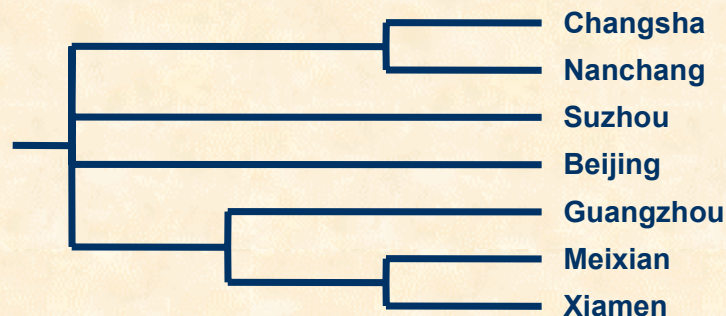
# Application to Chinese

- Word list of 100 meanings analyzed
- For each meaning (character), the forms that *appear* to be cognate are assigned the same character state

Index:	Meaning:		C	N	S	B	M	X	G
23	eat	吃	1	1	1	1	2	2	2
24	egg	蛋	1	1	1	3	2	2	1
25	eye	眼	1	1	1	1	2	2	1
26	feather	羽毛	1	2	2	1	2	1	2
33	give	给	1	1	2	3	4	5	2
35	grease	油脂	1	2	1	3	2	2	2
47	know	知道	1	1	1	2	2	2	2
72	say	说	1	3	2	2	1	1	1
78	small	小	1	2	2	2	1	1	1
80	stand	站	1	1	3	1	2	2	2
83	sun	太阳	1	2	1	1	2	2	3
84	swim	游泳	3	5	1	4	2	2	1
92	walk	走	1	1	1	1	2	2	2
95	what	什么	3	1	4	1	2	1	2
97	who	谁	1	1	2	3	5	2	4

character states for 15  
diagnostic meanings

Reconstructed consensus tree:



10 trees are equally parsimonious,  
each indicating 7 borrowings

# What can this tell us about vertical transmission?

- **Prune the most parsimonious trees based on other criteria** (Kitching, 1998):
  - the probability that observed sub-groupings would be observed in a random sample drawn (without replacement) from all possible binary tree topologies
  - the number of trees (having a binary feature) has Hypergeometric distribution

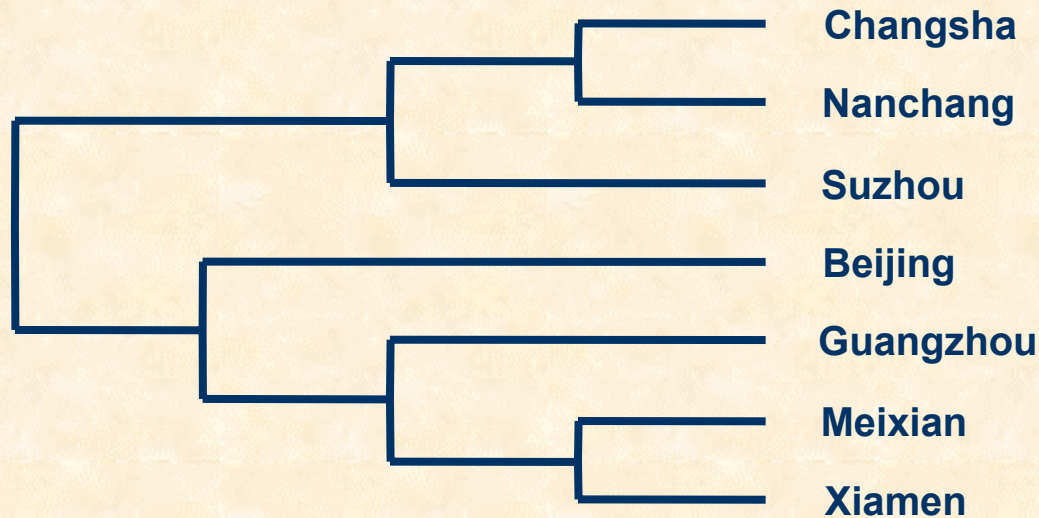
- **Test sub-grouping hypotheses, for example:**

(GZ, (MX, XM))	10 out of 10	$1.9 \times 10^{-16}$	<b>ACCEPT</b>
(BJ, (GZ, (MX, XM)))	3 out of 10	$1.1 \times 10^{-4}$	<b>ACCEPT</b>
(BJ, (SZ, (CS, NC)))	1 out of 10	$1.4 \times 10^{-2}$	—
((BJ, SZ), (CS, NC))	1 out of 10	$1.4 \times 10^{-2}$	—
(BJ, (CS, NC))	0 out of 10	1	—
(SZ, (CS, NC))	3 out of 10	$1.1 \times 10^{-4}$	<b>ACCEPT</b>
(CS, NC)	8 out of 10	$1.7 \times 10^{-7}$	<b>ACCEPT</b>

- **Collate trees that are consistent with accepted hypotheses**

# What can this tell us about vertical transmission?

- One consistent tree:



# What can this tell us about horizontal transmission?

- **In the 10 most parsimonious trees, borrowing is indicated for:**

“feather” (10)	“small” (10)	“what” (10)
“grease” (8)	“sun” (8)	“say” (6)
“know” (5)	“give” (2)	“who” (2)

- **In the 10395 possible trees, borrowing would be indicated for:**

“feather” (9900)	“small” (9900)	“what” (9180)
“grease” (9270)	“sun” (9792)	“say” (9270)
“know” (9900)	“give” (8610)	“who” (8610)

and for 6 other meanings

- **Of the 9 indicated borrowed meanings, which are significant?**

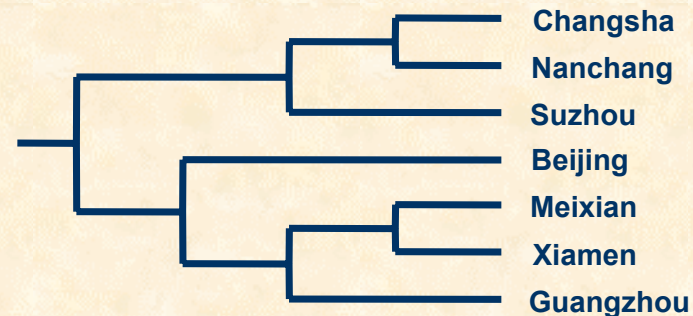
# What can this tell us about horizontal transmission?

- Hypothesis test: null hypothesis is that borrowing *has occurred*
- What are the probabilities that no more than these amounts of borrowing would occur by chance?

“feather” (1)	“small” (1)	“what” (1)
“grease” (.296)	“sun” (.111)	“say” (.017)
“know” (5e-5)	“give” (2e-5)	“who” (2e-5)

- At 1% significance, reject “know”, “give”, and “who” as borrowed

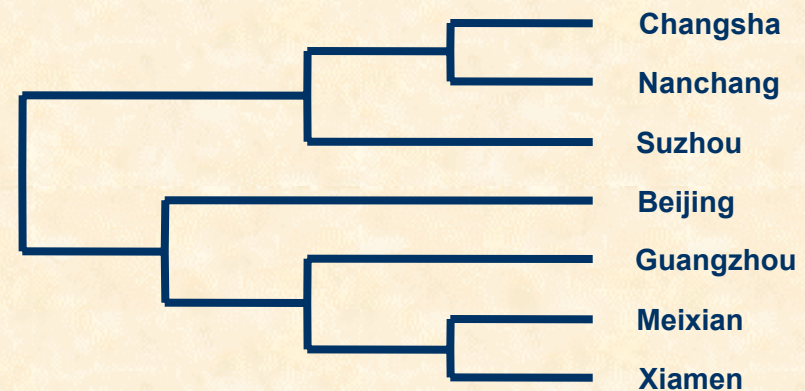
- The consensus tree of the five most parsimonious trees that reflect these is:



# What can this tell us about horizontal transmission?

- One of these 5 trees matches the tree that is consistent with the “vertical” analysis:

Meaning:		C	N	S	B	M	X	G
feather	羽毛	1	2	2	1	2	1	2
grease	油脂	1	2	1	3	2	2	2
say	说	1	3	2	2	1	1	1
small	小	1	2	2	2	1	1	1
sun	太阳	1	2	1	1	2	2	3
what	什么	3	1	4	1	2	1	2



- Borrowing indicated:**

“feather” (BJ, CS, XM × 2)

“what” (GZ, MX)

“sun” (MX/XM → NC)

“small” (GZ/MX/XM → CS or NC/SZ → BJ)

“grease” (GZ/MX/XM → NC or CS, SZ)

“say” (GZ/MX/XM → CS or SZ, BJ)

# Problems with the Algorithm

- **How to determine which character state is the retained state?**
  - currently, we use the “majority wins” principle  
i.e. the state that appears in the most taxa is considered the retained state  
but this is not always appropriate, e.g. a prestige language donor
  - **so ask a linguist!**  
but this is subjective
  - the algorithm should be extended to allow known retentions to be marked

# Conclusion

- **Can use cladistic analysis to perform genetic classification**
  - **Can distinguish *some* meanings for which words have been borrowed**
  - **Can determine sets of possible languages between which the borrowing took place**
  
  - **Should be extended to allow retained states to be specified (if known)**
  - **Should be extended to allow borrowing between language families to be detected**
- e.g. Northern (Huaxia) & Southern (Baiyue) Chinese dialects ?**

# References

- Baxter, W. H. (2000) “Beyond lumping and splitting: probabilistic issues in historical linguistics,” Time depth in historical linguistics, McDonald Institute for Archaeological Research, Cambridge.
- Bloomfield, L. (1933) *Language*, Allen & Unwin, London.
- Durie, M. & Ross, M. eds. (1996), *The comparative method reviewed: regularity and irregularity in sound change*, Oxford University Press, New York.
- Hennig, W. (1950) *Grundzüge einer Theorie der phylogenetischen Systematik*, Deutsche Zentralverlag, Berlin.
- Kitching, I. J. et al. (1998) *Cladistics: the theory and practice of parsimony analysis*, 2<sup>nd</sup> ed., Oxford University Press, New York.
- Ross, M. D. (1996) “Introduction,” in: Durie & Ross (1996).
- Ross, M. D. (1996b) “Contact induced change and the comparative method: cases from Papua New Guinea,” in: Durie & Ross (1996).
- Schleicher, A. (1861) *Compendium der vergleichenden Grammatik der indogermanischen Sprachen*, 3<sup>rd</sup> ed., Böhlau, Weimar.
- Schmidt, J. (1872) *Die Verwandtschaftsverhältnisse der indogermanischen Sprachen*, Weimar.
- Xu, T.-J. (1991) *Lishi Yuyanxue (Historical Linguistics)*, Shangwu Yinshuguan, Beijing.
- Zhou, Z.-H. & You, X.-L. (1986) “Fangyan yu Zhongguo Wenhua,” Renmin Chubanshe, Shanghai.