Work Papers of the Summer Institute of Linguistics, University of North Dakota Session

Volume 18 Article 8

1974

A computer analysis of Vietnam language relationships

Kenneth D. Smith
SIL-UND

A Computer Analysis of Vietnam Language Relationships *

Kenneth D. Smith

0. Introduction

David Thomas in his article "Mon-Khmer subgroupings in Vietman" (1966) presented the first lexicostatistical study of the Mon-Khmer languages in Vietnam. The cognate percentages so grouped themselves that the distinction between the Katuic and Bahnaric branches of the Mon-Khmer language family was very evident; that is, the relationships between the languages of one group and the languages of the other group were between 22-37%, whereas the relationships among the Katuic languages themselves and among the Bahnaric languages themselves ranged between 40-50%, except that closer groupings within both the Bahnaric and Katuic groups indicated an additional northern and southern division within each.

Four years later Thomas and Headley published a sequel: "More on Mon-Khmer subgroupings" (1970) which extended the lexicostatistical study to include Cambodian languages and establish a Pearic branch in Mon-Khmer (though distinct from the Khmer language which seems to form its own branch within Mon-Khmer). They also included Laotian languages and established a West-Bahnaric group within the Bahnaric branch. They further asserted that their computations "put Viet-Muong incontrovertably within the Austroasiatic phylum, and very likely within the Mon-Khmer family itself." In this latter study they retracted the north-south Katuic division proposed in the first paper.

Some of these classifications have been verified by phonological reconstructions, notably Proto-East Katuic (Thomas, 1967), Proto Mnong (Blood, 1968), Proto Bahnaric (Phillips, 1971), and Proto North Bahnaric (Smith, 1972).

---

* A paper presented to the Linguistic Forum of the Summer Institute of Linguistics (Norman, Oklahoma) on July 22, 1974

Only one aspect of the Thomas or Thomas and Headley classifications has come under attack. That relates to their inclusion of the Bahnar language within North Bahnaric. Though lexicostatistic evidence suggests a North Bahnaric classification, both Phillips (1971) and Smith (1972) in their phonological reconstructions suggested a South Bahnaric classification for Bahnar. The dispute was reviewed by Gregerson, Smith and Thomas (1973) in "The place of Bahnar within Bahnaric" which presented a compromise settlement by proposing a Central Bahnaric grouping for Bahnar and Alak (a language of southern Laos).

Meanwhile word lists from many other little known languages of Southeast Asia are becoming available and there is a great desire to establish their position within the language classification of other languages of the area. But the tedious and time-consuming nature of the task is overwhelming. Or, as Thomas (1973) has written:

> "For several years I have been doing Mon-Khmer lexicostatistical calculations, doing it all manually, both the comparing and the percentage calculation. If I had a dollar for every hour I have spent on it I would be a rich man. Much more needs to be done now, but only a computer would have the time and patience to do it. Also the cognate decisions need to be open to inspection and review. So I have decided that the time has come to computerize the project before I will look another word list in the face."

In the spring of 1974 while attending the University of Pennsylvania and taking a special Computational Linguistics course I had opportunity to write some computer programs and, being inspired by Thomas' letter and his general outline of program features, attempted to replace his worn patience with that of a computer. Furthermore, an attempt was made to investigate how results would differ if the basic 281-item word list used as input data was restricted, say, to the 100- or 2 0-word Swadesh list, or a specially compiled 212-word list--each of these being a subset of the entire 281-word list.

The programs described below were written in ALGOL-W.

1. Input data

The Summer Institute of Linguistics-Vietnam Branch has published word lists for most of the languages of south Vietnam including Khmer, Cambodian Cham, and Nung, a refugee group from North Vietnam (SIL 1968-1970). The word list for each language consists of an identical set of 281 items with glosses in English and Vietnamese. The 281-words include all words of both the Swadesh 100 and 200-word lists. Thirty-one such lists (a 32nd--Todrah--was added midway in the project) were used for this project. These include close dialects

(e.g. Chrau Jro and Chrau Prang, Hre BaTo and Hre SonHa, etc.) as well as the principal languages of the Katuic and Bahnaric branches of Mon-Khmer, Khmer, Vietnamese, and 8 Chamic languages (which, being Austronesian or MalayoPolynesian languages, are presumably
Fig.1) unrelated to the other Austroasiatic languages). See Figure 1 for a sample word list.

A 3x5 card was prepared for each of the 281 items, and headed with the word number, English and Vietnamese gloss. Each (tribal language) word of the first word list was then transferred to its respective card and assigned the number "1", representing the first word of cognate set #1. Then each word of the second word list (and subsequently all the word lists) was transferred to its respective card; if the word was judged cognate with a word already on the card, it was assigned to the same numbered cognate set, but if the word was judged non-cognate with a word already on the card, it was assigned to a new numbered cognate set. Some items of the word list had as few as one cognate set (#11); others as many as 22 cognate
Fig.2) sets (#262). See Figure 2 for a sample card showing several cognate sets for a given word of the basic word list. As each word of a given language word list was assigned a cognate set number, that number was written opposite that word on that word list. Note the hand-written numbers in Figure 1.

If two non-cognate words were cited for a given entry on a word list, each would be assigned a different cognate set number. But these two cognate set numbers are combined into a 3- or 4-digit number: the higher cognate set number being the last (or right) two digits (inserting a 0 for one-digit cognate set numbers) and the lower cognate set number being the first (or left) one or two digits. For example: 309 represents cognate set numbers 3 and 9, 1518 represents cognate set numbers 15 and 18, etc.

If three or more words are cited for a given entry on the word list they each were assigned a cognate set number on the card, but only two were selected for the written coding on the word list--usually those two which seemed most related to other languages.

If no tribal language word was cited for a given entry in the word list (i.e. the word list has a blank) then for that entry the cognate set number 0 was assigned.

Note that the linguist looks at each tribal language word on each word list only once, assigning it to one of an increasing number of cognate sets of words. He does not make a judgment for each pair of languages; the computer will do that utilizing the assigned cognate set numbers for each word of each word list.

For each language a set of 14 IBM cards was punched. The first card starts with the number 100 (identifying the beginning of a set of data for a given language) followed by the language

name (16 spaces) and a three-letter language name abbreviation. Each of the next 12 cards included in order the cognate set numbers for 25 consecutive words. The 25-word restriction per card enables quick location and correction or change of a cognate set number if desired. The 12th card had only the six cognate set numbers for words 276 through 281.

The last card had the number 101 identifying the end of a set of data for a given language.

2. Preliminary programs

2.1 Card check, program C1

Because of the amount of data included in this program--281 words for 32 languages or dialects, or 8,992 cognate set numbers--a preliminary program was written to check the input cards. The program did the following:

(1) read and printed the input parameters required for the main program;

(2) printed each language name and abbreviation;

(3) checked to see that each cognate set number was a realistic number, i.e. not less than 0 nor greater than the highest cognate set number established, or, for the 3- or 4-digit numbers, neither of the two conjoined cognate set numbers was greater than the highest cognate set number established;

(4) checked to see that each set of language data had exactly 281 cognate set numbers.

2.2 Data display, program C4(a)

The data display program provided a printout in one large chart of all the data. Columns are headed by the 3-letter language name abbreviation. Each of the 281 numbered lines gives the cognate set number assigned for each language, including the 3- or 4-digit conjoined cognate set numbers and "BLK" (blank) for each 0 cognate set number included. See Figure3. This program provides access to the data for inspection and review. (The Todrah data was subsequently printed out in Program C5 (b).

Fig. 3)

2.3 Widespread words, program C4(b)

For each of the 281 words, the computer searched for cognate sets which were represented in 16 or more of the 31 languages and dialects. It then printed a list citing the word number, cognate set number, and the number of languages or dialects whose words belong to the same cognate set. This produces an interesting list of the most widespread words of the area represented on the word lists.

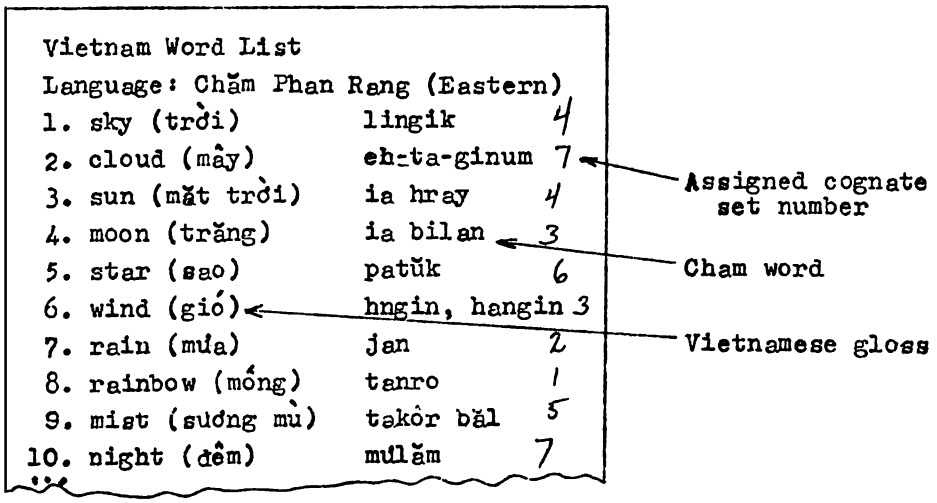


Figure 1. Portion of a word list

6. wind (gió)
   1 - kial, gió, khial, cayêu, kayal, chhal, syal
   2 - addiq
   3 - ngin, hngin/hangin
   4 - xeang
   5 - rabù
   6 - fă lŏm

Figure 2. Word card with cognate set numbers and forms

No.:	KMR	MNR	KTL	STG	PCH	CRJ	SRE	RGS	CME	JOR	BHP	RNG	SDG	SED	...
1	3	1	1	1	2	1	1	4	4	5	1	1	1	1	
2	6	2	2	2	5	2	1	1	7	710	1	2	2 3	2 3	
3	1	1	2	1	2	1 2	1 2		4	4	4	1	1	1	1
4	1	1	2	1	4	1	4	3	3	3	1	1	1	1	
5	5	3	4	3	4	3	3	6	6	6	1	1	1	1	
6	1	1	2	1	4	1	1	5	3	3	1	1	1	1	
7	5	1	4	1	4	1	1	2	2	2	1	1	1	1	
...															

Figure 3. Portion of full data display

	NUMBER OF WORDS---	COGNATE:	COMPARED:	% COGNATE:
VIETNAMESE	AND BRU	70	269	26
VIETNAMESE	AND PACOH	52	275	19
VIETNAMESE	AND HIGH KATU	50	266	19
VIETNAMESE	AND LOW KATU	48	266	18
VIETNAMESE	AND EASTERN CHAM	30	275	11
...				
BRU	AND PACOH	152	271	56
BRU	AND HIGH KATU	109	264	41
BRU	AND LOW KATU	107	264	41
BRU	AND EASTERN CHAM	35	267	13
...				

Figure 4. Portion of 465 two-language comparisons

3. Program for computation of cognate relationships

3.1 Computer program C2

The cognate percentage relationship of two languages was computed by comparing the cognate set numbers assigned to the word of each language for a given word of the word lists. If the two cognate set numbers are identical then the two languages share cognate forms for the particular word; if the two cognate set numbers are not identical then the two languages do not have cognate forms for that particular word. If one (or both) languages has a 3- or 4-digit number for a cognate set number code, representing two cognate sets, if either corresponds to the (or either) cognate set number of the other language, the two languages are considered to have cognates for that word. The computer thus counts the number of cognate pairs as well as the total number of pair compared (excluding any pair at least one of whose cognate set numbers is 0), divides the first by the second, multiplies by 100, and rounds off the result to obtain the percentage cognate relation between two languages.

The computer thus computes the percentage congate relation between every pair of two languages. For the complete 281 word list and 31 languages, the percentage cognate relation of 465 pairs of languages was computed, making 130,665 different 2-word comparisons.

Fig. 4 Figure 4 gives a portion of the prinout of the 465 two-language comparisons using the full 281-word list, stating number of cognates, number of words compared, and the percentage cognate figure.

Separate computation of the percentage cognate relationship was done using 4 different sets of words. For the shorter 100-, 200-, and 212-word lists, the computer skipped over those words not to be included.

3.2 Comparison of word lists

Fig. 5 Figure 5 compares the four sets of cognate percentages throughout the range of values measuring distance of all the languages from the Sedang languages. Those of the 212-word list are, on the average 1 or 2% higher than those of the 100-word list; and the 100-word list figures are an average of 2% higher than the 281-word list; and the 281-word list figures are an average of 2% higher than the 200-word list. The range from the lowest percentage to the highest percentage for a given pair of languages runs from 2% to 12%, with the lowest differences occurring for the language relationships of greater than 90% or lower than 30%; the greatest differences occur in the mid range of 30-90% cognate.

Figure 5 attempts to order the languages in an optimum ranking does not correspond to any of the individual orderings--each includes 2 or 3 misorderings in this optimally ordered composite

list. Consequently each of the four word lists appears relatively as valid as any other.

Comparing Thomas' (1966) figures with the four sets included here, his figures never exceed the bounds of the four figures computed here and most closely correspond to those of the 100-word list. (Thomas and Headley (1970.408) admitted to frequent differences of 5% on their respective judgment of cognateness, using a word list of 207 items, and suggest that their figures may be 5-10% lower than those in Thomas (1966).)

4. Program for structuring language relationships

4.1 Computer programs C3, C6, C7

A list of 465 percentages fails to reveal any structuring among the group of languages or indication of their genetic or historical development. In an attempt to use the computed cognate relationships to show such inter-relationships among these languages, the languages were ordered and arranged forming a triangular display with the percentage cognate figures occurring at the intersection of horizontal rows and vertical columns, each row or column headed by a language name. This triangular display lent itself readily to the preparation of the more common language tree (see below). This ordering was done separately for each of the four different word lists described above. The order of languages was determined as follows:

(1) The computer searched for the lowest percentage cognate figure for any two languages compared. These two languages, say A and Z, were assigned A to the leftmost column and Z to the bottom row.

(2) The computer then searched for (a) that remaining language (not Z) having the highest percentage cognate relation with A, say language B, as well as (b) that remaining language (not A) having the highest percentage cognate relation with z, say Y. Then, if the cognate percentage of A:B was greater than Y:Z, B was added to the triangle as the second column from the left (next to A), else, since Y:Z would be greater than A:B, Y was added as the next to bottom row (next to Z).

(3) Thereafter, that language which had the highest cognate relation with either the last added column or last added row would be added to the triangle adjacent to that column or row, respectively, until every language had been made part of the triangle. Figure 6 shows the triangle resulting from the cognate relationships using the 281-word list (lines drawn through the triangle are discussed below).

Fig. 6)

4.2 Language tree derivation

The triangular display of cognate percentages can be translated into a language tree diagram by the following procedures:

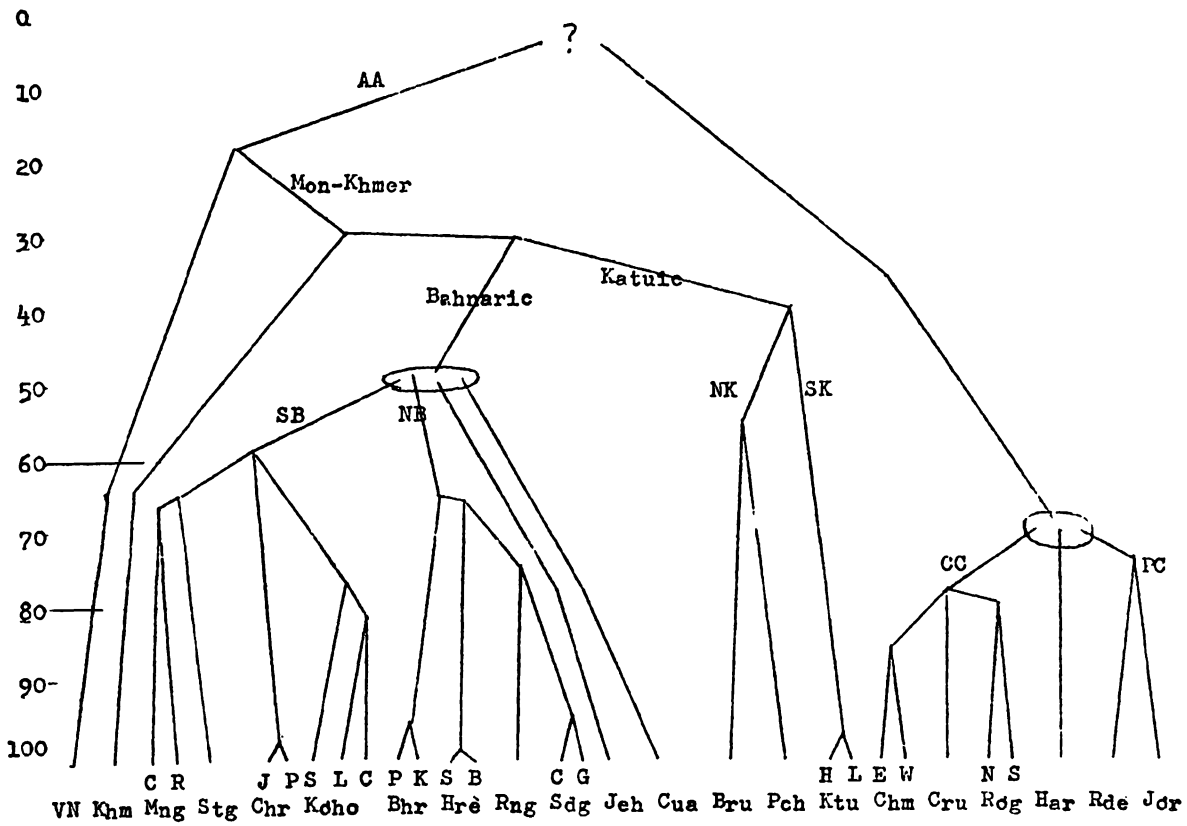
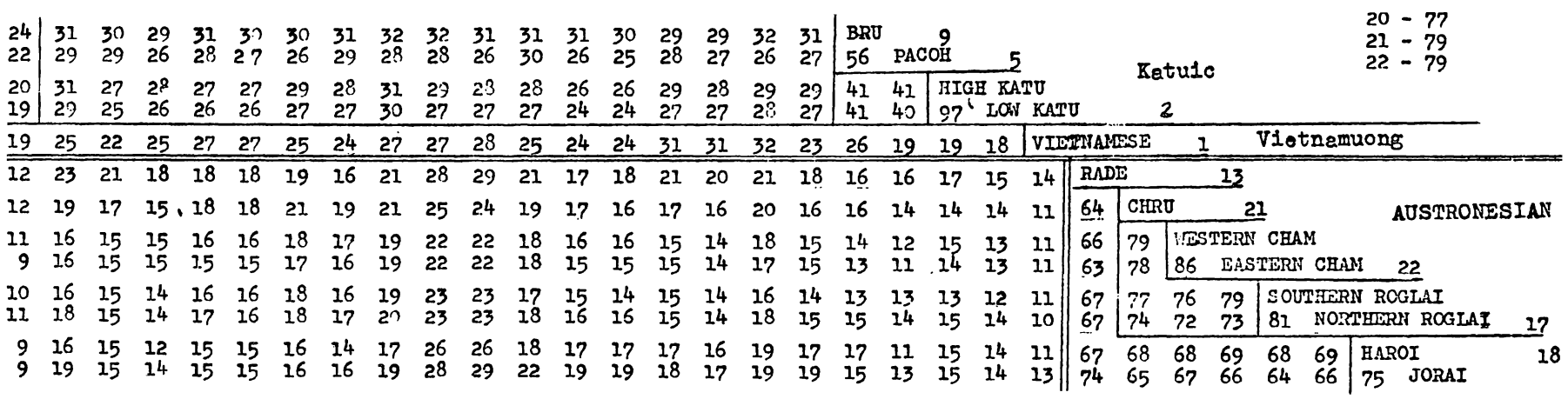


Figure 7. Language tree derived from 281-word list comparisons

Sedang:	212-word list	100-word list	281-word list	200-word list	difference	Thomas (1966)
Grtr Sdg	95	97	94	96	3	
Tơdrah	90	88	--	85	5	
Rengao	80	74	74	71	9	
Hrê BaTơ	72	76	66	68	10	
Hrê SơnHa	71	75	66	67	9	
Bahnar Ktm	67	64	62	59	8	
Bahnar Plk	65	63	59	57	8	
Jeh	67	58	60	56	11	
Cua	54	52	50	48	6	
Chrau Jro	49	47	44	39	10	44
Chrau Prang	49	47	44	39	10	
Mnong Rơlâm	49	46	44	39	10	46
Stieng	49	43	44	39	10	44
Koho Chil	49	40	44	39	10	
Koho Lach	48	38	44	39	9	
Mnong Central	50	42	43	38	12	
Koho Srê	47	39	42	36	11	39
Bru	31	33	30	28	5	31
High Katu	28	33	26	22	6	33
Low Katu	27	31	24	21	6	31
Vietnamese	26	30	24	26	10	
Pacoh	26	29	25	24	5	27
Khmer	27	23	25	24	4	
Jơrai	21	19	19	19	2	2
Radê	19	16	18	16	3	
Hơroi	18	16	17	16	2	
Chru	17	19	16	16	3	
N. Rơglai	17	18	16	15	3	
W. Cham	16	17	16	16	2	
E. Cham	15	17	15	14	3	
S. Rơglai	15	16	14	13	3	

Figure 5. Distance from Sedang using four different word lists and comparison with Thomas (1966)



(1) List in order from the lowest to the highest each percentage figure which occurs along the hypotenuse (i.e. at the top of a column or at the right end of a line). Number the list: 1, 2, 3.... (See right side of Fig. 6.)

(2) Starting at the point with the lowest percentage figure, or number 1 in (1) above, make successive divisions of the triangle into smaller triangles (by drawing lines between columns and rows), noting that the low percentage figure indicates a break between the languages listed above it and the languages listed to the right of it. Number each point where the division of the triangle is made, following the numbered sequence of (1) above.

(3) In drawing successive lines, do not cross a previously drawn line. The previously drawn line indicates a more major division of languages, whereas subsequently drawn lines indicate a division of languages within a smaller unit.

(4) Note that the percentage figures within the box formed below and to the left of the starting point at hypotenuse are generally lower percentage figures than those either above or to the right of the box. If the numbers should be larger in the box than those above or to the right of the box, stop the division of the triangle at that point. Such inversion of figures indicates an overlapping of language grouping relationships or a three-way division of languages.

(5) Each of the numbered points along the hypotenuse of the divided triangle is a node of a language tree, in order of chronological (cf. glottochronological) breaks. Successive numbered nodes may be in different branches of the tree, indicating only near simultaneous splits in different language areas. If a line was stopped, as suggested in (4) above, two numbered points may share a node of the language tree.

(6) Adjacent nodes representing close cognate percentages may not indicate exact order of language splitting.

Fig. 7) This scheme was used to prepare the language tree in Figure 7 from the percentages obtained from the 281-word list as given in Figure 6.

4.3 Comparison of language trees.

Comparing the language tree of Figure 7 with those from the 100-, 200-, and 212-word lists, the following similarities and differences are noted:

(1) All four trees distinguish very clearly:

(a) the divisions of Austroasiatic and Austronesian language stocks;

(b) the uniqueness of, and north-south division within, Katuic;

(c) the distant division of Bru and Pacoh within North Katuic;

(d) the division of Plateau Chamic and Coastal Chamic;

(e) the division of North and South Bahnaric except that the 100-word list places the Koho languages on an equivalent level;

(f) the remote association of Cua within Bahnaric or North

Bahnaric;

(g) that Bahnar is apart from Hre, Rengao, Sedang and Todrah.

(2) The four trees picture differently:

(a) the ordering of the remote breakoff of Khmer and Vietnamese from Austroasiatic (the 200-word list tree splits Katuic off before Vietnamese);

(b) the break off point of Jeh, whether before or within North Bahnaric;

(c) whether Haroi is Plateau Chamic or a third branch in addition to Plateau Chamic and Coastal Chamic.

These similarities and differences roughly coincide with the known and unknown aspects of language relationships resulting from comparative phonological studies.

5. Conclusion

To concur with and quote Thomas and Headley (1970.411) in summary:

> "To sum up, lexicostatistics is not a precision tool. Careful phonological reconstruction is necessary if one desires detailed information about language relationships. Lexicostatistics is useful, however, for giving a quick general picture of language groupings. Individual cognate percentages mean little, but clusterings of percentages can be meaningful and reliable, especially if separated by 5-10 percentage points from other clusterings...."

The above computer programs are therefore proposed as a means of determining language relationships without the tediousness of individual language comparisons but with the thoroughness and patient working that a computer offers us for processing great bulks of language data.

Bibliography

Blood, Henry F. 1968. A reconstruction of Proto Mnong. Grand Forks: SIL, UND. 115 p.; MA thesis, Indiana Univ. (1967).

Gudschinsky, Sarah C. 1956. "The ABCs of lexicostatistics (glottochronology)." Word 12.175-210; repub. in Dell Hymes, ed., Language in Culture and Society, 1964.

Gregerson, Kenneth J., Kenneth D. Smith, and David D. Thomas. 1973. "The place of Bahnar within Bahnaric." Paper presented at the First International Conference on Austroasiatic Linguistics, Honolulu, January 1973.

Phillips, Richard L. 1971. Phonological reconstruction of Proto-Bahnaric." PhD dissertation, Cornell Univ. In preparation.

Smith, Kenneth D. 1972. A phonological reconstruction of Proto North Bahnaric. Language Data, Asia Pacific series, No. 2, 106 p. SIL, Santa Ana.

SIL (Summer Institute of Linguistics). 1968, 1970. Vietnam word list. (A separate listing of 281 words for each of 32 languages.) Saigon. mi., 5 p. each

Thomas, David D. 1966. "Mon-Khmer subgroupings in Vietnam." In Norman Zide, ed., Studies in Comparative Austroasiatic Linguistics, p. 194-202.

_____. 1973. Personal letter to Marlene Laurence, SIL computer programmer in Saigon.

_____ and Robert K. Headley, Jr. 1970. "More on Mon-Khmer subgroupings." Lingua 25, 398-418.

Thomas, Dorothy. 1967. Proto East Katuic. Unpublished MA thesis, Univ. of North Dakota. vi, 103 p.