CHAPTER 4

LEXICOSTATISTICS

4.0 Introduction

To address the research question of the thesis, lexicostatistics is used to determine rough subgroupings of speech varieties.

Smith (1981), after comparing 281 words, provides percentages of lexical similarity as shown in Table 30:

KOHO 80-83%	Cil		
	83%	Lach	
	80%	81%	Sre

Table 30. Percentages of lexical similarity from Smith (1981)

Tạ Văn Thông (1988a) also compares the same 281 words as Smith (1981), and provides a chart of percentages of lexical similarity which can be seen in Table 31:

	Maa		Kodon		Sre-Nop		Cil	
Lach	221	79%	228	81.1%	233	82.9%	246	87.5%
Cil	225	80.1%	227	80.8%	232	82.6%		
Sre-Nop	254	90.4%	261	92.9%				
Kodon	261	92.9%						

Table 31. Percentages of lexical similarity from Tạ Văn Thông (1988a)

From this lexical comparison, Tạ Văn Thông divides the groups into two subgroups: Cil-Lach and Sre-Nop-Kodon-Maa. He further separates Sre-Nop from Maa-Kodon, because of geographical and phonetic distinctions. Finally he concludes that there are three main dialects: Northern (Cil-Lach), Central (Sre-Nop), and Southern (Maa-Kodon).

PAYAP UNIVERSITY

These two studies (Smith 1981 and Tạ Văn Thông 1988a) compare as shown in Table 32:

Cil		
83%/87.5%	**Lach**	
80%/82.6%	81%/82.9%	**Sre**

Table 32. Comparison of lexical percentages (Smith 1981/Tạ Văn Thông 1988a)

From these two studies, it can be seen that the same 281 words, collected and analyzed by different scholars, can yield results differing by 2-7%.

4.1 Lexicostatistics

This section provides background information for lexicostatistic study. It will discuss the lexicostatistic process and constructing tree diagrams. The section will also summarize the lexicostatistic results in the current study.

4.1.1 Wordlist Comparison

Wordlist comparison can be problematic since speech varieties have different forms and meanings. To have a good list of words used in lexicostatistic comparison, this thesis reviews some discussions of problems with wordlists.

In 1974, Huffman gave five reasons that several items in Swadesh's original list of 200 words are not applicable to Southeast Asian languages. These reasons are: 1) not culturally relevant or too abstract, 2) semantic mismatch, 3) compound words, 4) too general, and 5) loan words. Miller (1994) points out that none of the wordlists collected in national languages, such as Lao, Thai and Vietnamese have "completely similar semantic domains for all entries on the list" (1994:68). Some words are synonyms and homonyms. This lexicostatistic study uses 100 words selected by Mann (2002). The chosen lexical items should help to minimize some of these

PAYAP UNIVERSITY

problems by using the most common items from Southeast Asian comparative wordlists.

The purpose of the lexical analysis in this thesis is to determine the linguistic relationship between Koho and Maa varieties. The closeness of the varieties is based on the degree of lexical similarity, that is, by the number of apparent cognate forms found in the varieties.

4.1.2 Lexicostatistics Process

Koho and Maa word forms tend to be sesquisyllabic. Sesquisyllabic word forms are a syllable and a half in length. Sesquisyllabic languages have a presyllable which comes first and tends to be unstressed while the following main syllable is stressed. These minor syllables are subject to change over time and may collapse onto the main syllable leaving only remnants of the presyllable or may be lost entirely. Thomas (1992:206) explains the sesquisyllabic structure as,

> ...a result of stress shift to the final syllable of the word (stress group), followed by increasing phonetic nucleation at the final syllable. This progressive strengthening of the final syllable leads to progressive weakening of the non-final syllables, so that the penultimate syllable in a disyllabic word tends toward fusion with the final syllable.

Since lexical comparison attempts to determine which forms are cognate across the varieties, it is not appropriate to compare presyllables. The cognate forms here places more weight on the main syllables. For instance, supposing there are two words with and without the presyllable, if the main syllables are similar, these words are considered as cognates despite the absence or presence of the presyllable. From the Maa and Koho wordlists, it can be seen that only a few of the varieties have preserved the presyllables. Therefore, according to this lexicostatistic method, if presyllables are not ignored, it may underestimate the lexical similarity percentage. The elimination of the presyllable forms is illustrated in the comparison shown in Table 33 which gives

PAYAP UNIVERSITY

the full syllable forms, including the presyllable, and Table 34 which provides the root forms after the presyllable has been removed:

Ref. No.	**362**	**159**	**090**	**043**	**315**	**044**	**004**	**214**
English	**black**	**bone**	**tail**	**leaf**	**to kill**	**flower**	**star**	**smoke**
Vietnamese	**đen**	**xương**	**đuôi**	**lá**	**giết**	**hoa**	**sao**	**khói**
Maa Dagui	ʔo:t	nti:ŋ	tiəŋ	nha:	srəp	ka:w	səmaɲ	ɲhuʔ
Maa Chop	ʔo:ɲ	nti:ŋ	tieŋ	nha	rəsət	ka:w	sərmaɲ	ɲhuʔ
Maa Tadung	ɟu:ʔ	nti:ŋ	ntiəŋ	nha	rəchət	bəka:w	sərmaɲ	ɲhuʔ
Koho Cil	ɟu:ʔ	rti:ŋ	tiɛŋ	nha:	təncət	bəka:w	səmaɲ	ɲhoʔ
Koho Lach	ɟuʔ	kəntiŋ	tieŋ	ləha:	tənchət	pəka:w	sərmaɲ	ɲhoʔ
Koho Nop	ɟu:ʔ	kati:ŋ	nchan	nha:	kəsət	bəka:w	səmaɲ	ɲhuʔ
Koho Sre	ɟu:ʔ	nti:ŋ	ntiaŋ	nha:	kəchət	bəka:w	səmaɲ	ɲhuʔ

Table 33. Full syllable forms

Ref. No.	**362**	**159**	**090**	**043**	**0315**	**044**	**004**	**214**
English	**black**	**bone**	**tail**	**leaf**	**to kill**	**flower**	**star**	**smoke**
Vietnamese	**đen**	**xương**	**đuôi**	**lá**	**giết**	**hoa**	**sao**	**khói**
Maa Dagui	ʔo:t	ti:ŋ	tiəŋ	ha:	rəp	ka:w	maɲ	huʔ
Maa Chop	ʔo:ɲ	ti:ŋ	tieŋ	ha	sət	ka:w	maɲ	huʔ
Maa Tadung	ɟu:ʔ	ti:ŋ	tiəŋ	ha	chət	ka:w	maɲ	huʔ
Koho Cil	ɟu:ʔ	ti:ŋ	tiɛŋ	ha:	cət	ka:w	maɲ	hoʔ
Koho Lach	ɟuʔ	tiŋ	tieŋ	ha:	chət	ka:w	maɲ	hoʔ
Koho Nop	ɟu:ʔ	ti:ŋ	chan	ha:	sət	ka:w	maɲ	huʔ
Koho Sre	ɟu:ʔ	ti:ŋ	tiaŋ	ha:	chət	ka:w	maɲ	huʔ

Table 34. Main syllable or “root” forms

The words selected for comparison were also scrutinized to eliminate possible morphological markers, and non-root syllables are ignored. Using the data collected, the relationship between word forms is fairly clear. Pairs of roots are then compared on a phoneme-by-phoneme basis, with each pair of phonemes assigned to one of three categories somewhat similar to a method described in Blair (1990:31-32) to clarify which forms appear to be cognate and which forms are not cognate. The following are the criteria used in the lexicostatistic comparison for this thesis:

PAYAP UNIVERSITY

Category	**Criteria**
Cat. 1	a. exact matches b. vowels differing by one feature (includes diphthongs) c. vowels differing by length d. regular sound correspondences
Cat. 2	a. phonetically similar consonants b. vowels differing by two or more features (includes length and diphthongs)
Cat. 3	a. non-phonetically similar consonants b. non-regularly occurring deletions
Ignore	a. pitch distinctions b. presyllables c. supplemental semantic morphemes d. regularly occurring deletions e. reduplicated syllables

Category 1(d) and 2(a) relate to the condition of phonetically similar segments. To meet the requirement of 1(d) the recurrence must be greater than that required to satisfy the criterion of 2(a). Category 1(b) and 2(b) consider vowels or diphthongs differing by one feature have a higher degree of similarity than those differing by two or more features.

The lexical similarity of word forms is determined based on these criteria. This yields the following result shown in Table 35:

PAYAP UNIVERSITY

Ref. No.	362	159	90	43	315	44	004	214
English	black	bone	tail	leaf	to kill	flower	star	smoke
Vietnamese	đen	xương	đuôi	lá	giết	hoa	sao	khói
MDg-MCh	1a, 3a	1a, 1a, 1a	1a, 1b, 1a	1a, 1a	3a, 1a, 2a	1a, 1a, 1a	1a, 1a, 1a	1a, 1a, 1a
MDg-MTd	3b, 1b, 2a	1a, 1a, 1a	1a, 1a, 1a	1a, 1a	3a, 3b, 1a, 2a	1a, 1a, 1a	1a, 1a, 1a	1a, 1a, 1a
MDg-KhCl	3b, 1b, 2a	1a, 1a, 1a	1a, 1b, 1a	1a, 1a	3a, 1a, 2a	1a, 1a, 1a	1a, 1a, 1a	1a, 1b, 1a
MDg-KhLch	3b, 1b, 2a	1a, 1c, 1a	1a, 1b, 1a	1a, 1a	3a, 3b, 1a, 2a	1a, 1a, 1a	1a, 1a, 1a	1a, 1b, 1a
MDg-KhNp	3b, 1b, 2a	1a, 1c, 1a	3a, 3b, 2b, 2a	1a, 1a	3a, 1a, 2a	1a, 1a, 1a	1a, 1a, 1a	1a, 1a, 1a
MDg-KhSr	3b, 1b, 2a	1a, 1c, 1a	1a, 1b, 1a	1a, 1a	3a, 3b, 1a, 2a	1a, 1a, 1a	1a, 1a, 1a	1a, 1a, 1a
MCh-MTd	3b, 1b, 3a	1a, 1a, 1a	1a, 1b, 1a	1a, 1a	3a, 3b, 1a, 1a	1a, 1a, 1a	1a, 1a, 1a	1a, 1a, 1a
MCh-KhCl	3b, 1b, 3a	1a, 1a, 1a	1a, 1b, 1a	1a, 1a	3a, 1a, 1a	1a, 1a, 1a	1a, 1a, 1a	1a, 1b, 1a
MCh-KhLch	3b, 1b, 3a	1a, 1c, 1a	1a, 1a, 1a	1a, 1a	3a, 3b, 1a, 1a	1a, 1a, 1a	1a, 1a, 1a	1a, 1b, 1a
MCh-KhNp	3b, 1b, 3a	1a, 1c, 1a	3a, 3b, 2b, 2a	1a, 1a	1a, 1a, 1a,	1a, 1a, 1a	1a, 1a, 1a	1a, 1a, 1a
MCh-KhSr	3b, 1b, 3a	1a, 1c, 1a	1a, 1b, 1a	1a, 1a	3a, 3b, 1a, 1a	1a, 1a, 1a	1a, 1a, 1a	1a, 1a, 1a
MTd-KhCl	2a, 1c, 1a	1a, 1a, 1a	1a, 1b, 1a	1a, 1a	1a, Ig, 1a, 1a	1a, 1a, 1a	1a, 1a, 1a	1a, 1b, 1a
MTd-KhLch	2a, 1a, 1a	1a, 1c, 1a	1a, 1b, 1a	1a, 1a	1a, 1a, 1a, 1a	1a, 1a, 1a	1a, 1a, 1a	1a, 1b, 1a
MTd-KhNp	2a, 1c, 1a	1a, 1c, 1a	3a, 3b, 2b, 2a	1a, 1a	3a, 3b, 1a, 1a	1a, 1a, 1a	1a, 1a, 1a	1a, 1a, 1a
MTd-KhSr	1a, 1c, 1a	1a, 1c, 1a	1a, 1b, 1a	1a, 1a	1a, 1a, 1a, 1a	1a, 1a, 1a	1a, 1a, 1a	1a, 1a, 1a
KhCl-KhLch	1a, 1a, 1a	1a, 1c, 1a	1a, 1b, 1a	1a, 1a	1a, Ig, 1a, 1a	1a, 1a, 1a	1a, 1a, 1a	1a, 1a, 1a
KhCl-KhNp	2a, 1c, 1a	1a, 1c, 1a	3a, 3b, 2b, 2a	1a, 1a	3a, 1a, 1a	1a, 1a, 1a	1a, 1a, 1a	1a, 1b, 1a
KhCl-KhSr	1a, 1a, 1a	1a, 1c, 1a	1a, 1b, 1a	1a, 1a	1a, Ig, 1a, 1a	1a, 1a, 1a	1a, 1a, 1a	1a, 1b, 1a
KhLch-KhNp	2a, 1c, 1a	1a, 1a, 1a	3a, 3b, 2b, 2a	1a, 1a	3a, 3b, 1a, 1a	1a, 1a, 1a	1a, 1a, 1a	1a, 1b, 1a
KhLch-Sr	1a, 1c, 1a	1a, 1a, 1a	1a, 1b, 1a	1a, 1a	1a, 1a, 1a ,1a	1a, 1a, 1a	1a, 1a, 1a	1a, 1b, 1a
KhNp-KhSr	2a, 1a, 1a	1a, 1a, 1a	3a, 3b, 2b, 2a	1a, 1a	3a, 3b, 1a, 1a	1a, 1a, 1a	1a, 1a, 1a	1a, 1a, 1a

Table 35. Lexical similarity analysis

Key: MDg = Maa Dagui, MCh = Maa Chop, MTd = Maa Tadung, KhCl = Koho Cil, KhLch = Koho Lach, KhNp = Koho Nop, and KhSr = Koho Sre

Once the categories have been assigned for all of the phones, a similarity matrix is used to determine whether the words are lexically similar or not. This matrix is based on the phone length of the words compared and specifies the minimum conditions the word forms must meet in order to be considered lexically similar. The phone table for minimum lexical similarity is shown in Table 36.

PAYAP UNIVERSITY

Phones		Category 1	Category 2	Category 3
1	=	1	0	0
2	=	2	0	0
3	=	2	1	0
4	=	2	1	1
5	=	3	1	1
6	=	3	2	1
7	=	4	2	1
8	=	4	2	2

Table 36. Phone table for minimum lexical similarity (Blair 1990:32)

Using this phone table, most of the words in seven varieties are lexically similar. The words 'black' and 'to kill' in Maa Dagui, however, are not cognate with the other speech varieties. Likewise, the word 'black' in Maa Chop does not appear to resemble to the other varieties either. The word 'tail' in Koho Nop is not lexically similar to other varieties. Table 37 illustrates which items are considered lexically similar.

PAYAP UNIVERSITY

Ref. No. English Vietnamese	362 black đen	159 bone xương	90 tail đuôi	43 leaf lá	315 to kill giết	44 flower hoa	004 star sao	214 smoke khói
MDg-MCh	Not sim	Lex sim	Lex sim	Lex sim	Not sim	Lex sim	Lex sim	Lex sim
MDg-MTd	Not sim	Lex sim	Lex sim	Lex sim	Not sim	Lex sim	Lex sim	Lex sim
MDg-KhCl	Not sim	Lex sim	Lex sim	Lex sim	Not sim	Lex sim	Lex sim	Lex sim
MDg-KhLch	Not sim	Lex sim	Lex sim	Lex sim	Not sim	Lex sim	Lex sim	Lex sim
MDg-KhNp	Not sim	Lex sim	Not sim	Lex sim	Not sim	Lex sim	Lex sim	Lex sim
MDg-KhSr	Not sim	Lex sim	Lex sim	Lex sim	Not sim	Lex sim	Lex sim	Lex sim
MCh-MTd	Not sim	Lex sim	Lex sim	Lex sim	Not sim	Lex sim	Lex sim	Lex sim
MCh-KhCl	Not sim	Lex sim	Lex sim	Lex sim	Not sim	Lex sim	Lex sim	Lex sim
MCh-KhLch	Not sim	Lex sim	Lex sim	Lex sim	Not sim	Lex sim	Lex sim	Lex sim
MCh-KhNp	Not sim	Lex sim	Not sim	Lex sim	Lex sim	Lex sim	Lex sim	Lex sim
MCh-KhSr	Not sim	Lex sim	Lex sim	Lex sim	Not sim	Lex sim	Lex sim	Lex sim
MTd-KhCl	Lex sim	Lex sim	Lex sim	Lex sim	Lex sim	Lex sim	Lex sim	Lex sim
MTd-KhLch	Lex sim	Lex sim	Lex sim	Lex sim	Lex sim	Lex sim	Lex sim	Lex sim
MTd-KhNp	Lex sim	Lex sim	Not sim	Lex sim	Not sim	Lex sim	Lex sim	Lex sim
MTd-KhSr	Lex sim	Lex sim	Lex sim	Lex sim	Lex sim	Lex sim	Lex sim	Lex sim
KhCl-KhLch	Lex sim	Lex sim	Lex sim	Lex sim	Lex sim	Lex sim	Lex sim	Lex sim
KhCl-KhNp	Lex sim	Lex sim	Not sim	Lex sim	Not sim	Lex sim	Lex sim	Lex sim
KhCl-KhSr	Lex sim	Lex sim	Lex sim	Lex sim	Lex sim	Lex sim	Lex sim	Lex sim
KhLch-KhNp	Lex sim	Lex sim	Not sim	Lex sim	Not sim	Lex sim	Lex sim	Lex sim
KhLch-Sr	Lex sim	Lex sim	Lex sim	Lex sim	Lex sim	Lex sim	Lex sim	Lex sim
KhNp-KhSr	Lex sim	Lex sim	Not sim	Lex sim	Not sim	Lex sim	Lex sim	Lex sim

Table 37. Lexical similarity

Key: MDg = Maa Dagui, MCh = Maa Chop, MTd = Maa Tadung, KhCl = Koho Cil, KhLch = Koho Lach, KhNp = Koho Nop, KhSr = Koho Sre, Lex = Lexical, and sim = similarity

Finally, this data is summed up. The same process for a hundred words is applied with the resulting table expressed as a percent of lexical similarity for the lexical items examined:

PAYAP UNIVERSITY

Koho Cil						
82%	**Koho Lach**					
79%	80%	**Koho Sre**				
78%	77%	90%	**Maa Tadung**			
78%	81%	94%	91%	**Maa Chop**		
77%	76%	87%	89%	91%	**Maa Dagui**	
70%	74%	89%	85%	85%	79%	**Koho Nop**

Table 38. A matrix of lexical similarity percentages

Percentages are put in a matrix which shows the lowest figures on the lower left side, while the figures generally increase as moving away from the lower left. From the cognate percentages of each pair of languages, a language tree diagram may be constructed.

4.1.3 Tree Diagram

A tree diagram of the Koho and Maa varieties under lexicostatistic investigation is shown in Figure 9. The tree diagram was generated by the "Unweighed Pairs Grouped Method with Arithmetic Average" (UPGMA, or Average Link) method. The table of apparent cognate percentages was processed using two computer programs, "Cluster Analysis 1.01" (Quigley 1995), and also the "Neighbor" program (part of the PHYLIP 3.6 suite of programs, Felsenstein 2002), which also implements the UPGMA method. Figure 9 shows the rooted (UPGMA) tree produced by the Cluster Analysis 1.01 program.[5]

From Figure 9, it can be seen that Koho Cil and Koho Lach stand together in one group, while the other five speech varieties stand together in another grouping. Interestingly, it appears that "Koho" is not a cohesive grouping. "Maa" may be a somewhat cohesive group.

PAYAP UNIVERSITY

[5] I gratefully acknowledge the assistance provided by Dr. J. F. Bennett who both ran and explained these programs to me.

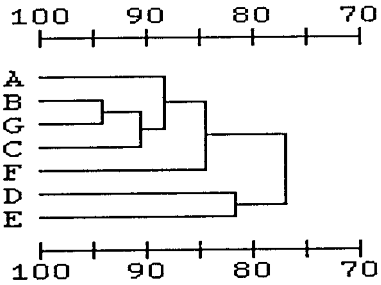

Figure 9. Rooted tree of Koho and Maa varieties based on UPGMA method

Key: A = Maa Dagui, B = Maa Chop, C = Maa Tadung, D = Koho Cil, E = Koho Lach, F = Koho Nop, G = Koho Sre.

Percentages of lexicostatistic similarity show that Koho Cil and Koho Lach group together, with the other five varieties grouped in one branch. Based on the tree above, it is obvious to see that Maa varieties are cohesive, but Koho ones are not. It is worth to note that the Koho Sre, which has been traditionally grouped with Koho Lach and Koho Cil, is actually more closely related to the Maa speech varieties. Koho Nop is less similar to the Maa varieties (and Koho Sre) than the other Maa varieties are to each other; however, it is still more similar to Maa than to Koho Cil and Koho Lach. This result leads to a conclusion that these Koho and Maa speech varieties are geographically grouped, since the Koho Lach and Koho Cil inhabit the area in the north and the rest settle in the south of the Province as shown in Figure 10.

PAYAP UNIVERSITY

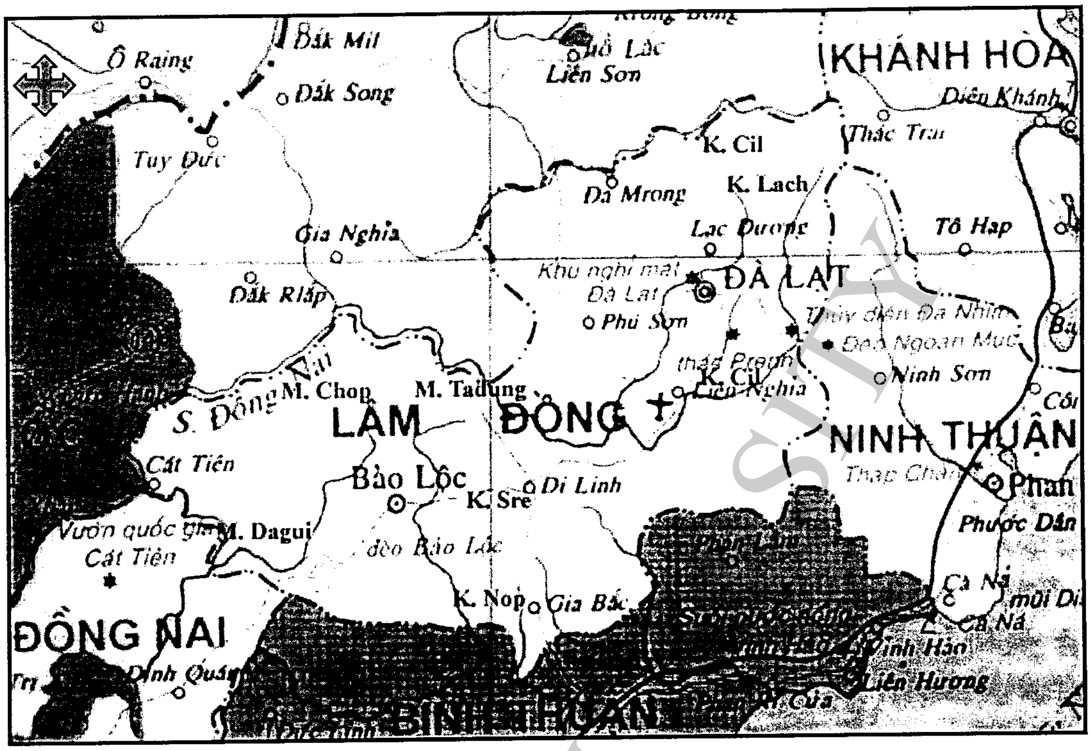

Figure 10. Location of the Koho and Maa

The map in Figure 10 shows the locations of Koho and Maa speech varieties. The Koho Lach and Koho Cil are located with respect to all Maa varieties, Koho Sre, and Koho Nop.

4.2 Summary

This chapter covered lexicostatistic considerations of Koho and Maa speech varieties. Using lexical similarity percentages, tree diagrams were created. The study found that the Maa grouping is more cohesive than the Koho grouping. Koho Cil and Koho Lach tend to be in one group, while the other Koho varieties tend to group more closely with the Maa varieties. These groupings tend to make good sense from a geographical perspective since Koho Cil and Koho Lach have settled in the north of Lam Dong

PAYAP UNIVERSITY

Province while the Maa group together with Koho Sre and Koho Nop who inhabit an area to the south of the Province.

This lexicostatistic study provides us a preliminary grouping from which emerges a general picture of the speech varieties. Nonetheless, a comparative phonological reconstruction provides a more reliable picture of the language relationships as Thomas and Headley (1970:411) state:

> 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 next chapter provides a detailed phonological reconstruction to capture the relationships of Koho and Maa speech varieties.

PAYAP UNIVERSITY