Bulletin of the Institute of History and Philology
Volume 61, Part 4 (December 1990)
Date Published : December 1992

Classification of Formosan Languages: Lexical Evidence

Paul Jen-kuei Li

The purpose of this paper is to present some new lexical evidence for the classification of Formosan languages, as based on all Formosan cognates that have been identified. Fourteen extant and five extinct Formosan languages are compared in this study. Based on the number of cognates shared by each pair of languages, the distance between each pair of languages is calculated. A tree diagram for Formosan languages is then constructed by adopting the procedure developed for quantitative studies by Cavalli-Sforza and Edwards (1967), and Fitch and Margoliash (1967). Since the number of lexical items (about 400) compared for each of the five extinct languages is much smaller than the number (about 1,000) compared for the other fourteen extant languages, some adjustment has been made with regard to the number of cognates shared by each pair of languages by adopting Jaccard's coefficient principle. Furthermore, cognates shared exclusively by a pair of languages are regarded as providing much stronger evidence for a close genetic relationship. The results of this study roughly agree with classifications for these languages based on other types of evidence, including phonological and syntactic evidence. Slightly different results are found when each language is represented by only one dialect, as opposed to being represented by more than one dialect. All these results are more suggestive than conclusive.

1. Introduction

The purpose of this paper is to present some new lexical evidence for the classification of Formosan languages.[1] We shall show a way of quantifying the degree of lexical similarities among a group of languages. We believe that lexical evidence can be regarded as one

1 This manuscript was prepared while I was visiting the Department of Linguistics, University of California, Berkeley, 1989--1990. I am indebted to Lien-meng Lin at the Academia Sinica for writing computer programs for Tables 1-2, 6 and to Zhong-wei Shen at UC Berkeley for Tables 3-5 as well as for showing me how to measure distances between languages as indicated in the figures in this paper. I have profited from valuable comments and suggestions for improvement from William S-Y Wang, Stanley Starosta, Shigeru Tsuchida, Pang-hsin Ting, Zhong-wei Shen and Kathleen Ahrens. They may not agree with all that I have presented here. I remain responsible for any errors that remain in this paper.

piece of evidence for language subgrouping. Moreover, we feel that subgrouping languages on the basis of lexical evidence is as valid as hypothesizing a genetic relationship on the basis of phonological and/or syntactic evidence.

It is assumed in this paper that the degree of lexical similarity between languages, i.e. shared cognates determined on the basis of regular sound correspondences, is somehow correlated with the degree of genetic relatedness. Although there is no guarantee that the correlation always holds true, we may still quantify genetic relatedness with this working hypothesis, and then compare results obtained by this method with those obtained by other methods, such as the standard comparative method, if available.

The method adopted in this paper is different from traditional lexicostatistics in several respects. First, only a pair of languages is compared at a time in lexicostatistics, whereas in our procedure all languages are compared at the same time. Hundreds of typologically possible trees will be examined and only one best possible tree will be chosen by the computer. Thus errors can be detected. How well the final chosen tree fits the input data can be checked by the statistic value --standard deviations. Second, traditional lexicostatistics uses only a 100 or 200 word basic vocabulary, whereas for our comparison we utilize all vocabulary that is available, which may amount to hundreds or thousands of lexical items. The so called "basic vocabulary" may vary from language to language, and the decision between "basic" and "non-basic" is often arbitrary. Moreover, we can minimize errors if we deal with a large amount of data.

2. Previous Classifications of Formosan Languages

Classification of Formosan languages has long been an issue, and there is no one classification that is generally accepted by Formosan scholars. In the past quarter of the century, Formosan languages have generally been divided into three subgroups : (1)the Atayalic, including Atayal and Sediq, (2)the Tsouic, including Tsou, Kanakanavu and Saaroa, and (3)the Paiwanic, including all the rest; see Dyen (1963, 1971b), and Ferrell (1969, 1972, 1979a, b). Tsuchida (1976) formally proposed to group Rukai with the Tsouic languages,

and that was further supported by Dyen (1987a, b). After examining fourteen phonological and syntactic features among all living Formosan languages, Ho (1983) argued that Rukai had a closer genetic relationship with the Paiwanic rather than with the Tsouic languages. Li (1985) proposed to group Saisiyat, Pazeh and four northwestern extinct languages (Taokas, Babuza, Papora and Hoanya) with the Atayalic languages, based on lexical and phonological as well as syntactic evidence.

Evidence for close relationships between some languages is obvious, and so has generally been accepted. For instance, no one has disputed the close genetic relationship between Atayal and Sediq and that between Kanakanavu and Saaroa. Despite the fact that there is strong lexical and phonological evidence for the close relationship between Tsou and the two Southern Tsouic languages (Tsuchida 1976), proposing a genetic relationship for these three languages on the basis of their morphological and syntactic evidence is less certain (Ferrell 1972). Based on evidence in Tsou's morphological and syntactic evolution, Starosta (1985) argues for the unique position of Tsou by itself. [2] However, Tsuchida (personal communication, Sept. 1990) believes that he can demonstrate that Tsou is morphologically very close to Kanakanavu and Saaroa. Moreover, morphologically Rukai is the most unique among all Formosan languages.

3. The Data

Ever since July 1970, I have been doing extensive field work on virtually all the living Formosan languages: Atayal, Sediq, Tsou, Kanakanavu, Saaroa, Rukai, Bunun, Puyuma, Thao, Saisiyat, Pazeh, Kavalan and Amis. Paiwan data are based on Ho (1978) and his field notes. I have collected data for more than one dialect for most of these lan-

2 As of today, Starosta (personal communication, Nov. 1989) still believes that "Tsou is profoundly different from other Formosan languages in morphology (the focus system, the absence of *-AN in the verbal morphology and the total absence of reflexes of *-EN) and in syntax (the pervasiveness of the auxiliaries). These two facts could be connected, and maybe all the differences can be explained in terms of the extension of the auxiliaries and the final segment loss, but I found some paradoxes that I think could not be explained in that way alone." However, Tsuchida (personal communication) points out that *-AN and *-EN have been merged as *-*a*> -*a*. Also Kavalan has no trace of *-EN.

guages. I utilized about 1,000 lexical items for this comparative work.

The following Formosan languages are extinct: Taokas, Babuza, Papora, Hoanya and Siraya. There are only about 400 lexical items available for these languages, as listed in Tsuchida (1982). The qualities of these language data are uneven. They were recorded by various different people, trained or untrained in linguistics, and data for the "same" language may have been collected from different villages or speech communities. Some were transcribed in phonetic symbols; others were written in Chinese characters or in Japanese *katakana*. A variety of lexical forms may be listed for the same lexical entry, and they may or may not indicate different dialects or varieties of speech.

There are even less data available for the other extinct Formosan languages: Ketagalan, Basay and possibly Kulon (Tsuchida 1985). Thus they have been excluded from this comparative study.

4. The Cognates and the Genetic Relationships between the Languages

I have tried to identify Formosan cognates, internal as well as external to Formosan, off and on over these years. I can generally identify more cognates, on a more solid basis, for the languages that I have worked with longer and thus have become better acquainted with, such as Atayal, Sediq, Rukai, Thao, Saisiyat, Pazeh and Kavalan. I have included and updated the list of cognates that have been identified by other Formosan scholars such as Tsuchida (1976, 1982, 1985) and Dahl (1981). As shown on the diagonal in Table 1 below, the number of cognates that have been identified for Atayal is 396, Sediq 359, Tsou 300, Kanakanavu 379, Saaroa 374, Rukai 423, Bunun 232, Paiwan 329, Puyuma 265, Thao 207, Saisiyat 267, Pazeh 227, Kavalan 182, Amis 258, Taokas 85, Babuza 99, Papora 83, Hoanya 83, and Siraya 95. The total number of cognates that have been identified for all Formosan languages is 1,109.[3]

In Tables 1 & 2 below, each Formosan language is generally represented by a major or

3 In the roughly 1,000 lexical entries, many of them show no cognacy among Formosan languages. However, there may be more than one cognate set for the same lexical item.

important dialect: Atayal by Mayrinax, Sediq by Tongan, Tsou by Duhtu, Rukai by Budai, Bunun by Takituduh, Paiwan by Butanglu, Puyuma by Pinan, Saisiyat by Taai, Pazeh by Pazeh (not Kahabu), and Amis by Sakizaya.[4] If cognates are missing or unavailable for these major dialects, then they may be cited from some other dialects.

The number of cognates shared between Atayal and Sediq is 318, between Atayal and Tsou 88, between Sediq and Tsou 77, and so on. Generally speaking, the more closely related each pair of languages is, the more cognates they share with each other, as expected. The smallest number of cognates shared by a pair of extant Formosan languages is 67 between Sediq and Puyuma, and shared by an extant and an extinct language is 20 between Sediq and Hoanya, and between Tsou and Taokas.

Nonetheless, a large number of cognates shared by a pair of languages does not necessarily always indicate that these two languages have a close relationship. They may happen to retain more cognates that have been inherited from their common ancestor or parent language, such as Proto-Austronesian (PAN), Proto-Hesperonesian (PHN), and Proto-Formosan (PFN).[5]

However, cognates that are shared exclusively by a pair of languages are much stronger evidence for a close relationship between the two. All such cognates are much more likely lexical innovations than lexical retentions. If they were retentions, they would be shared by more than two languages, Formosan or extra-Formosan. I have gone through the list of cognates that are exclusively shared by each pair of languages, and none of them go back to the higher level nodes (such as PAN, PFN) or even lower level nodes (such as

4 Abbreviations for the language names are: Ata, Atayal; Sed, Sediq; Tso, Tsou; Kan, Kanakanavu; Sar, Saaroa; Ruk, Rukai; Bun, Bunun; Pai, Paiwan; Puy, Puyuma; Tha, Thao; Sai, Saisiyat; Paz, Pazeh; Kav, Kavalan; Ami, Amis; Tao, Taokas; Bab, Babuza; Pap, Papora; Hoa, Hoanya; Sir, Siraya.

5 Ting Pang-hsin (personal communication) believes that a large number of cognates shared by languages should indicate a close relationship, especially if the cognates go back to an early stage, just as shared phonological retentions from an early stage indicate a close relationship between languages. This position can be supported with abundant evidence from Chinese written documents. For instance, among all major Chinese dialects, only Min dialects still retain vocabulary such as *tia* (鼎) 'pan,' *lang* (儂) 'person,' which can be traced back to very early Chinese written documents. Geographically adjacent languages or dialects tend to share more common vocabulary, as in Cantonese, Hakka, Min and Wu dialects, and are apt to borrow from each other throughout history.

Table 1. Number of Cognates Shared by Each Pair of Languages

	Ata	Sed	Tso	Kan	Sar	Ruk	Bun	Pai	Puy	Tha	Sai	Paz	Kav	Ami	Tao	Bab	Pap	Hoa	Sir
Ata	396																		
Sed	318	359																	
Tso	88	77	300																
Kan	94	86	217	379															
Sar	94	85	219	299	374														
Ruk	100	92	180	233	232	423													
Bun	103	92	120	124	126	125	232												
Pai	89	83	140	165	167	216	120	329											
Puy	76	67	114	115	125	161	99	175	265										
Tha	90	86	100	102	101	109	124	111	97	207									
Sai	137	110	114	116	121	126	121	126	108	120	267								
Paz	116	90	87	92	88	95	91	97	89	106	138	227							
Kav	79	70	94	95	92	111	86	103	95	85	99	79	182						
Ami	98	92	114	128	127	143	123	151	140	118	128	105	125	258					
Tao	28	24	20	24	25	24	29	27	28	33	39	40	26	35	85				
Bab	31	30	36	37	40	42	40	46	46	47	43	47	33	47	48	99			
Pap	31	28	29	30	31	34	30	39	37	38	36	42	29	45	36	48	83		
Hoa	23	20	30	34	34	32	33	34	33	37	29	36	30	38	37	49	52	83	
Sir	37	31	52	62	61	61	49	59	55	53	50	38	45	58	27	39	32	35	95

PRT) in the derivation history. With few exceptions, they all belong to the lowest level node for each pair of languages.

Phonological innovations exclusively shared by two or more languages have been taken as evidence for a close genetic relationship between the languages by historical linguists. Similarly, cognates shared exclusively by a pair of languages can also be taken as indicating a close genetic relationship. [6] All cognates in Formosan languages are determined on the basis that they must observe the rules of regular sound correspondences, which have first been worked out by Tsuchida (1976) and subsequently revised and expanded by Li (1985). While many of these cognates go back to an early stage in the parent language, such as PAN, PHN, and PFN, many others go back to the lower level nodes such as PSF, PNF, PAt, PRT, PT, PR and PNW. [7] Only relatively few cognates are exclusively shared by a pair of languages. If a pair of languages exclusively shares a fair amount of cognates with each other, that means that they must have shared a long history of common development, not only in their phonology but also in their lexicon.

In Table 2, a pair of languages that has been known to have a closer relationship by the comparative method or simply by inspection does show a higher number of exclusively shared cognates, for instance, 180 cognates between Atayal and Sediq, 25 between Kanakanavu and Saaroa, 19 between Paiwan and Puyuma, 19 between Paiwan and Rukai, and so on. Yet between many pairs of languages, there are no or very few exclusively shared cognates, for instance, none between Atayal and Kanakanavu, Saaroa, Rukai, Puyuma, Thao or Amis, and only 1 between Atayal and Tsou, Paiwan or Kavalan.

Does the large number of cognates exclusively shared by Rukai and the Tsouic languages (9 between Tsou and Rukai, 17 between Kanakanavu and Rukai, 11 between

6 Dyen (1987b) has applied this method, which he calls "homomerous method," to subclassifying related languages such as Indo-European and Formosan languages with some interesting results.

7 Of all the 1,109 Formosan cognates that have been identified in this study, there are 174 PAN cognates, 163 PHN cognates, 124 PFN cognates, 213 Proto-Southern-Formosan (PSF) cognates, 29 Proto-Northern-Formosan (PNF) cognates, 175 Proto-Atayalic (PAt) cognates, 72 Proto-Rukai-Tsouic (PRT) cognates, 30 Proto-Tsouic cognates, 22 Proto-Rukai (PR) cognates, 35 Proto-Northwestern cognates, and so on. Proto-Northern-Formosan refers to the northern Formosan subgroup, including Atayal, Sediq, Saisiyat, Pazeh, and four northwestern extinct languages (see Li 1985), while Proto-Southern-Formosan refers to the Tsouic and the Paiwanic languages (cf. Tsuchida 1976).

Table 2. Number of Cognates Exclusively Shared by Each Pair of Languages

	Ata	Sed	Tso	Kan	Sar	Ruk	Bun	Pai	Puy	Tha	Sai	Paz	Kav	Ami	Tao	Bab	Pap	Hoa	Sir
Ata	197																		
Sed	180	181																	
Tso	1	0	31																
Kan	0	0	7	54															
Sar	0	0	8	25	49														
Ruk	0	1	9	17	11	67													
Bun	2	2	1	1	1	1	22												
Pai	1	0	2	2	1	19	2	53											
Puy	0	0	0	0	1	9	2	19	44										
Tha	0	0	1	0	0	0	4	1	0	12									
Sai	9	1	2	0	1	2	0	0	1	1	31								
Paz	4	1	1	0	0	1	2	2	1	2	12	33							
Kav	1	0	0	0	0	0	1	0	2	1	0	1	13						
Ami	0	0	0	1	0	1	4	3	7	3	2	1	2	25					
Tao	0	0	0	1	0	0	0	0	1	0	2	3	0	1	18				
Bab	0	0	0	0	0	0	0	0	0	0	1	1	0	0	9	16			
Pap	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	3	11		
Hoa	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2	3	6	12	
Sir	0	0	0	0	0	0	1	1	1	0	0	1	0	0	0	0	0	0	3

Saaroa and Rukai) indicate that they have a close genetic relationship, as Tsuchida and Dyen have argued? Our lexical evidence seems to lend support to such a hypothesis. The close relationship between Kanakanavu and Saaroa is self-evident at all levels of grammar, just as our lexical evidence is strong. Yet, our lexical evidence for the relationship between Tsou and these two Southern Tsouic languages seems rather weak: Only 7 exclusively shared cognates between Tsou and Kanakanavu, 8 cognates between Tsou and Saaroa, as compared with 9 cognates between Tsou and Rukai. This type of lexical evidence may suggest that Tsou be unique among all Formosan languages, as Starosta (1985) has suggested in his study of Tsou morphology and syntax. Furthermore, it may also suggest that Kanakanavu and Saaroa should be grouped with Rukai in the Paiwanic subgroup. If so, then the controversy over the position of Rukai in its relations with the "Tsouic" languages can easily be resolved.

The above results may have been somewhat skewed (or twisted) by the methods of treatment and calculation. Kanakanavu and Saaroa, which are very closely related, are treated as separate languages. However, Rukai has three divergent "dialect" groups that are not mutually intelligible to each other: (1)Budai, Tanan and Labuan, (2)Maga and Tona, and (3)Mantauran; see Li 1977. Yet they are treated as if they were the "same" language. When a certain cognate is not retained in a dialect group, it may be kept in another. Thus Rukai tends to share larger numbers of cognates with many other languages. Further study is required to separate Rukai into three groups to see if any of them still share a large number of cognates with any of the "Tsouic" languages.

Geographical adjacency is another factor that affects the number of cognates shared by each pair of languages. Kavalan has had close contact with Amis only in the past one hundred years or so, yet they share 125 cognates with each other, a number higher than the average, which is only 82.57. Puyuma has been geographically adjacent to Amis for a much longer period of time, and they share 140 cognates with each other, considerably higher than the average. Thao has been surrounded and heavily influenced by Bunun (see Li 1978), and they share 124 cognates. Rukai has been bordered by Paiwan for an unknown period of time, and they share 216 cognates with each other. As a matter of fact,

based on a comparison of lexical, phonological and syntactic similarities between Rukai and Paiwan, Ho (1983) has argued that Rukai is closer to the Paiwanic languages rather than to the Tsouic languages; cf. Fig. 3 in Section 7 below. Rukai has also been geographically located close to the Tsouic languages, especially the Southern Tsouic languages. For instance, as based on my own observations, there must have been a lot of borrowing that took place between the Mantauran dialect of Rukai and Saaroa of the Tsouic. It is sometimes difficult to separate between historical inheritance and borrowing, especially in an early stage before sound changes take place.

Since our lexical data for the five extinct languages are more limited: some 400 items as compared with 1,000 items for the other Formosan languages, we expect to get smaller numbers of cognates for these extinct languages. That is precisely what has turned out to be the case. Nonetheless, even with the restricted data available, we still find some impressive results: The number of exclusively shared cognates between Taokas and Babuza is 8 and that between Papora and Hoanya is 6, as shown in Table 2.

5. The Procedure for Drawing a Family Tree [8]

Table 3 shows the distance between each pair of Formosan languages. [9] There is a shorter distance between a pair of languages which share a larger number of cognates and are thus more closely related to each other. Conversely, there is a longer distance between a pair of languages which share a smaller number of cognates and are thus more distantly related.

Since the number of lexical items compared for each of the 5 extinct languages is not the same as the ones compared for the other 14 living Formosan languages, we have to make some adjustment with regard to the number of cognates shared by each pair of languages. We have done so by adopting Jaccard's coefficient principle (see below).

8 I am grateful to Zhong-wei Shen for showing me how to go about the whole procedure, which has been developed for quantitative studies by Cavalli-Sforza and Edwards (1967), Fitch and Margoliash (1967).

9 In Tables 3-5 and Figure 1 below, the Formosan languages are listed in the same order as in Tables 1 and 2 above. Thus 1 stands for Atayal, 2 for Sediq, 3 for Tsou, and so on.

Cognates are a type of linguistic character that can occur as one of two states: present (+)or absent (–). If one considers any two languages, x and y, the data for all the identified cognates can be summarized in a 2 x 2 table of counts having the following form:

		Lg. x	
		+	–
Lg. y	+	a	b
	–	c	d

$p = a + b + c + d$

P equals the total number of cognates identified in all 19 languages; **a** is the number of cognates present in both languages **x** and **y**; **b** is the number of cognates present in language **y** but absent in language **x**; **c** is the reverse; and **d** is the number of cognates absent in both languages.

Many similarity coefficients have been proposed for binary data of this type (see Clifford and Stephenson 1975). One that has commonly been used in numerical taxonomy is Jaccard's coefficient. This coefficient is defined as follows:

$$S_{x,y} = \frac{a}{a + b + c}$$

It is the ratio **r** of the number of positive matches **a** to the total number of cognates p minus the number of negative matches **d**. This coefficient ranges from zero, when there is no shared cognate (**a**=0), to unity, when two languages share all cognates with each other (**b**=0,**c**=0).

As mentioned above, the number of lexical items compared for each of the 5 extinct languages is not the same as the ones compared for the other 14 living languages. Since Jaccard's coefficient excludes the cognates which are absent in both languages, the coefficient of a pair of languages will not be very much affected by the total number of cognates identified in all languages.

Let us take Atayal and Sediq as an illustration. The total number of Atayal cognates is 396, among them 318 are shared with Sediq and the rest, 78 in all, are shared with some other languages. According to the 2 x 2 table above, 318 is **a**, 78 is **b**, and 396 = **a** + **b**. The total number of Sediq cognates is 359, the number of cognates shared with Atayal is the same, 318, and the rest of its cognates, 41, are shared with languages other than Atayal. So,

Table 3. Distance between Each Pair of Languages

19 L																		
1																		
2	.272																	
3	.855	.868																
4	.862	.868	.530															
5	.861	.869	.519	.341														
6	.861	.867	.669	.591	.589													
7	.804	.816	.709	.745	.737	.764												
8	.860	.863	.714	.696	.688	.597	.728											
9	.870	.880	.747	.783	.757	.694	.751	.582										
10	.825	.821	.754	.789	.790	.791	.606	.739	.741									
11	.740	.787	.748	.781	.767	.777	.680	.732	.745	.661								
12	.771	.819	.802	.821	.828	.829	.753	.789	.779	.677	.612							
13	.842	.851	.758	.796	.802	.775	.738	.748	.730	.720	.717	.761						
14	.824	.825	.743	.749	.749	.734	.665	.654	.634	.660	.678	.724	.603					
15	.938	.943	.945	.945	.942	.950	.899	.930	.913	.873	.875	.853	.892	.886				
16	.933	.930	.901	.916	.908	.912	.863	.880	.855	.819	.867	.832	.867	.848	.647			
17	.931	.932	.918	.931	.927	.928	.895	.895	.881	.849	.885	.843	.877	.848	.727	.642		
18	.950	.953	.915	.921	.920	.932	.883	.910	.895	.854	.910	.869	.872	.875	.718	.632	.544	
19	.919	.927	.848	.850	.850	.867	.824	.838	.820	.787	.840	.866	.806	.803	.824	.748	.781	.755
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
	Ata	Sed	Tso	Kan	Sar	Ruk	Bun	Pai	Puy	Tha	Sai	Paz	Kav	Ami	Tao	Bab	Pap	Hoa

for Sediq 318 is **a**, 41 is **b**, and 359 = **a** + **c**. Thus the coefficient can be calculated by the following formula :

$$S_{x,y} = \frac{\text{shared cog.}}{\text{total cog. in Lg.x} + \text{total cog. in Lg.y} - \text{shared cog.}}$$

The coefficient between Atayal and Sediq, therefore, is:

$$S_{x,y} = \frac{a}{(a+b)+(a+c)-a} = \frac{318}{396+359-318} = 0.7276$$

The distance (D) between Atayal and Sediq is:

$$D = 1 - r = 1 - 0.7276 = 0.272$$

Following this procedure, we can get the distance between all the pairs of Formosan languages, as given in Table 3, which is in essence based on Table 1.

Table 3 is used as an input matrix for our tree-drawing program. Table 4 is an output matrix of the tree-drawing program, which minimized the summation of squared errors. [10] Based on the figures in Table 4A, the best possible tree will be constructed.

Table 5 gives the length of each branch. Numbers 1 to 19, representing the 19 Formosan languages, are terminal nodes. The numbers above 19, i.e. between 20 and 36, are non-terminal nodes.

---

10 According to the "tree hypothesis," language evolution is a successive branching process. Let's consider four languages, a, b, c and d. There are fifteen ways in which these languages may have arisen from a single proto-language by successive binary splits. The number of different trees increases very rapidly with the number of languages. It is reported by Cavalli-Sforza and Edwards (1967) that $(2t-3)! / [2^{t-2}(t-2)!]$ different rooted trees can be reconstructed for t (= the number of) languages. When t = 10, this equals 34,459,425 trees. They also state that $(2t-5)! / [2^{t-3}(t-3)!]$ different unrooted trees are possible. When t = 10, this equals 2,027,025 trees. In practice, unless there are very few languages to be studied, a thorough examination of all the possible trees is far too complex.
One way to reduce the number of possible trees is to minimize the number of mutation steps (Fitch and Margoliash 1967). Thus a tree can be obtained by minimizing the difference between the distance among the languages in a constructed tree and the distance among the languages in the original data.

$$\text{Sum of squares} = \sum_{i=1}^{j} \frac{(\text{obs} - \text{exp})^2}{\text{obs}^2}$$

obs(erved) = the original distance
exp(ected) = distance in a constructed tree

Table 4. Distances between Each Pair of Languages for a Tree Diagram

A. Distance between languages on this tree

2	0.27																	
3	0.87	0.88																
4	0.88	0.89	0.53															
5	0.87	0.89	0.52	0.34														
6	0.86	0.88	0.61	0.62	0.61													
7	0.80	0.81	0.76	0.77	0.76	0.75												
8	0.83	0.84	0.69	0.70	0.70	0.69	0.72											
9	0.85	0.86	0.71	0.72	0.71	0.70	0.74	0.58										
10	0.80	0.81	0.76	0.76	0.76	0.75	0.61	0.72	0.73									
11	0.76	0.77	0.77	0.78	0.77	0.76	0.70	0.73	0.75	0.70								
12	0.79	0.80	0.80	0.81	0.80	0.79	0.73	0.76	0.78	0.73	0.61							
13	0.83	0.85	0.77	0.78	0.77	0.76	0.73	0.73	0.74	0.72	0.73	0.76						
14	0.79	0.80	0.72	0.73	0.73	0.72	0.68	0.68	0.70	0.68	0.69	0.72	0.60					
15	0.97	0.98	0.94	0.95	0.95	0.94	0.87	0.90	0.92	0.87	0.87	0.90	0.91	0.86				
16	0.93	0.94	0.90	0.91	0.91	0.90	0.83	0.86	0.88	0.83	0.83	0.86	0.87	0.82	0.68			
17	0.95	0.96	0.93	0.93	0.93	0.92	0.86	0.89	0.90	0.85	0.85	0.88	0.89	0.85	0.70	0.64		
18	0.96	0.97	0.93	0.94	0.93	0.92	0.86	0.89	0.91	0.86	0.85	0.88	0.89	0.85	0.70	0.64	0.54	
19	0.90	0.91	0.87	0.88	0.88	0.86	0.80	0.83	0.85	0.80	0.80	0.83	0.84	0.79	0.79	0.75	0.78	0.78
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
	Ata	Sed	Tso	Kan	Sar	Ruk	Bun	Pai	Puy	Tha	Sai	Paz	Kav	Ami	Tao	Bab	Pap	Hon

B. Error matrix

2	0.00																	
3	-0.02	-0.02																
4	-0.02	-0.02	0.00															
5	-0.01	-0.02	-0.00	0.00														
6	-0.00	-0.01	0.06	-0.02	-0.02													
7	0.00	0.00	-0.05	-0.02	-0.03	0.01												
8	0.03	0.02	0.02	-0.00	-0.01	-0.09	0.01											
9	0.02	0.02	0.04	0.07	0.04	-0.01	0.01	0.00										
10	0.03	0.01	-0.00	0.02	0.03	0.04	0.00	0.02	0.01									
11	-0.02	0.02	-0.02	0.00	-0.01	0.01	-0.02	0.00	-0.00	-0.04								
12	-0.02	0.02	0.00	0.01	0.03	0.04	0.02	0.03	0.00	0.05	0.00							
13	0.01	0.00	-0.01	0.02	0.03	0.02	0.01	0.02	-0.01	0.00	-0.02	-0.00						
14	0.03	0.02	0.02	0.02	0.02	0.02	-0.02	-0.03	-0.07	-0.02	-0.01	0.00	0.00					
15	-0.03	-0.04	0.00	-0.01	-0.00	0.01	0.02	0.03	-0.01	0.00	0.01	-0.04	-0.01	0.02				
16	0.00	-0.01	-0.00	0.01	0.00	0.02	0.03	0.02	-0.02	-0.01	0.04	-0.02	0.00	0.03	-0.03			
17	-0.02	-0.03	-0.01	-0.00	-0.00	0.01	0.04	0.01	-0.02	-0.00	0.03	-0.04	-0.01	0.00	0.03	0.01		
18	-0.01	-0.01	-0.01	-0.02	-0.01	0.01	0.02	0.02	-0.01	-0.00	0.06	-0.01	-0.02	0.03	0.01	-0.01	0.00	
19	0.02	0.02	-0.02	-0.03	-0.03	0.00	0.02	0.01	-0.03	-0.01	0.04	0.04	-0.03	0.01	0.03	-0.01	0.00	-0.02
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
	Ata	Sed	Tso	Kan	Sar	Ruk	Bun	Pai	Puy	Tha	Sai	Paz	Kav	Ami	Tao	Bab	Pap	Hon

Table 5. Length of Each Branch

Betweem	And	Length
1	20	0.13030
20	2	0.14170
20	32	0.29915
32	28	0.01935
28	24	0.00774
24	27	0.04104
27	10	0.30034
27	7	0.30566
24	30	0.01317
30	31	0.04213
31	14	0.27988
31	13	0.32312
30	23	0.03562
23	25	0.05861
25	6	0.30024
25	21	0.04857
21	22	0.09422
22	5	0.16810
22	4	0.17290
21	3	0.25972
23	26	0.04382
26	9	0.29949
26	8	0.28251
28	36	0.08799
36	19	0.36124
36	34	0.07282
34	33	0.01320
33	35	0.05897
35	18	0.27337
35	17	0.27063
33	16	0.30593
34	15	0.35890
32	29	0.03694
29	12	0.32016
29	11	0.29184

Figure 1 shows the structure of an unrooted tree, indicating language relationships. The branches in this tree diagram are not proportional; they simply indicate relationships between nodes. That is to say, the tree simply represents relationships of the languages. It does not indicate historical derivations or directions of language change. The computer

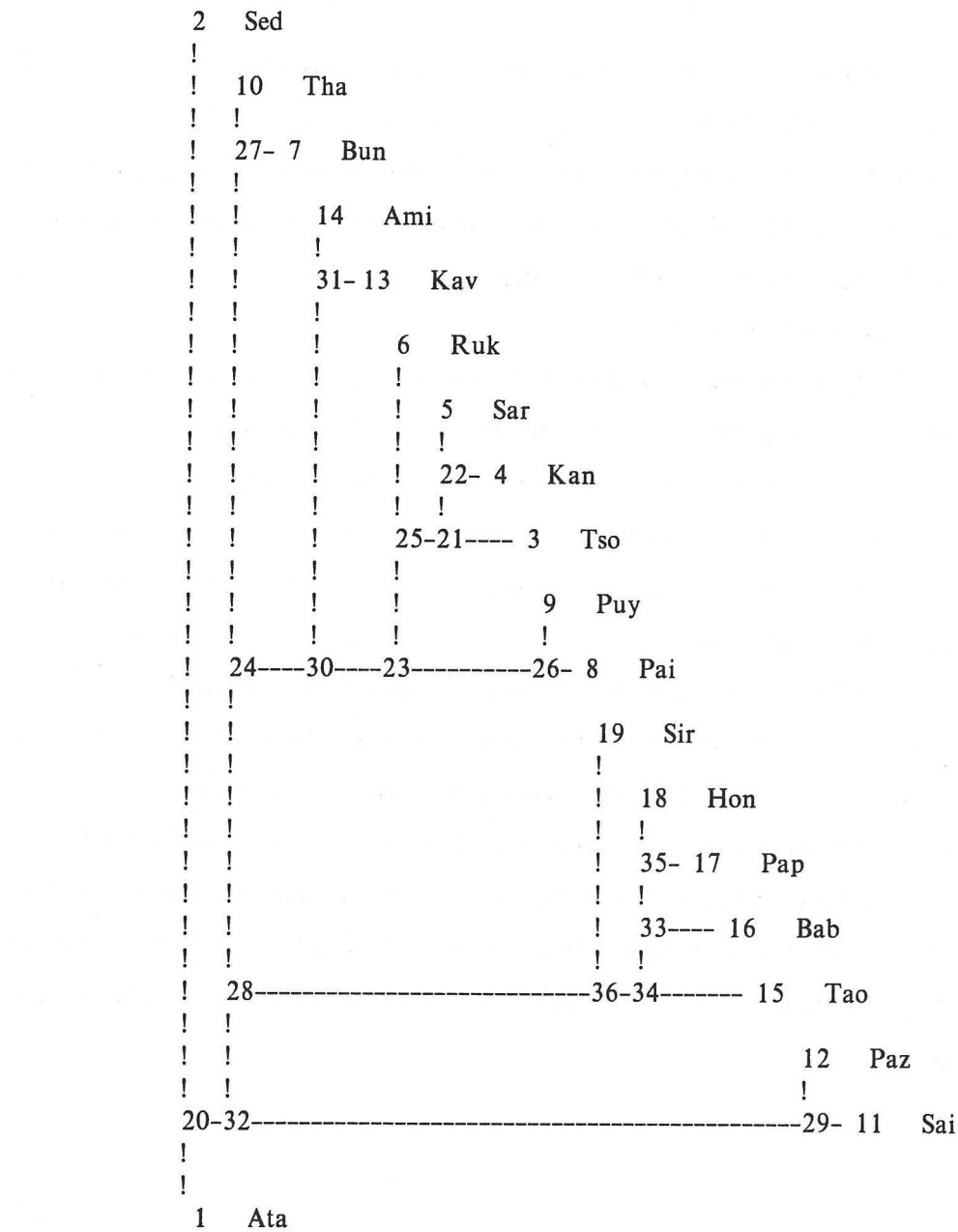
program has examined 601 trees and picked the most appropriate one as shown in Figure 1.

From Table 4 we can see that the largest distance between two languages is 0.98, which is the distance between Language 2 (Sediq) and Language 15 (Taokas). The midpoint between the two language nodes farthest apart, i.e. between Sediq and Taokas, can be chosen as the root of the tree, by which we mean the proto-language of all the 19 Formosan languages compared. After obtaining the tree, the directions of historical change for these languages can be determined.

Let it be noted that each derivation in the tree is always binary. This implies that at each stage of a language split, it can only split into two, and not more than two, branches. Although this is only a claim, it is more likely that language actually splits into two, but not more than two at any point in history. That is to say, binary splits, rather than tertiary splits, are perhaps closer to reality in language history. A computational tree may show that a pair of languages share either a relatively long or a short period of common history. In the case of the former, the results are more reliable, whereas in the latter they are shakier, due to our limited data. When more updated data become available, we may get a somewhat different tree indicating somewhat different genetic relationships.

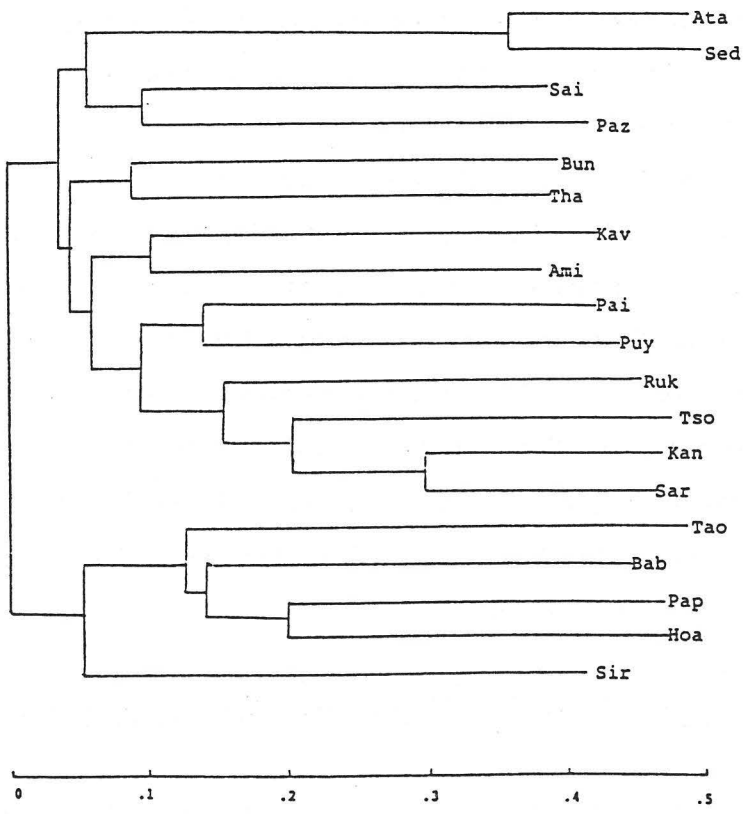
Based on Table 5, which indicates the length between every two connected nodes, and Figure 1, from which the root of the tree is located, a tree diagram of the Formosan languages under study can be drawn, as shown in Figure 2. Note that the relative length of each vertical line leading from each node indicates the time depth of each split for these languages.

Figure 1. Structure of Unrooted Trees Showing Relationships



Remember: this is an unrooted tree!
Sum of squares = 0.32283
Average percent standard deviation = 3.08142
601 trees examined

Figure 2. Formosan Classification (Each language is represented by one or more than one dialects)



6. A Family Tree for Formosan Languages

As indicated in Figure 2, our lexical evidence for the close relationship among the four northwestern languages (Taokas, Babuza, Papora and Hoanya) lends further support to the phonological evidence given by Tsuchida (1982:9-11). Similarly, it also lends support to the hypothesis that Saisiyat and Pazeh have a close relationship among themselves, and also have a relatively close relationship with the Atayalic languages, as suggested in Li

(1985). Yet it seems to come as a surprise that Siraya and the four northwestern languages are shown to constitute one of the two major subgroups of Formosan languages. The limited data available for these extinct languages can account for the low percentage of cognates that can be identified for these languages. We are limited not only by the data available, but also have little control over the quality. [11] Our results, with regard to the unique position of the five extinct Formosan languages, can at best be considered only a suggestion. They should not be taken at face value.

The problem of inter-dialectal borrowing cannot be easily resolved. As shown in Figure 2, both Saisiyat and Pazeh have a closer relationship with the Atayalic languages, and similarly the close relationships between Paiwan and Puyuma, between Kavalan and Amis, and between Bunun and Thao are not very certain. The close relationships in these groups of languages could all be attributed to borrowing, as each of these four groups of languages has been geographically close (see Tsuchida 1983). Some of the borrowing must have taken place recently: Kavalan has been heavily influenced by Amis, and so has Thao by Bunun. [12] As mentioned above, Rukai and the Tsouic languages may also have influenced each other for a long time.

7. Another Family Tree for Formosan Languages

The foregoing classification of Formosan languages is based on the cognate sets, in which some languages (including Kanakanavu, Sarroa, Thao and Kavalan) are represented by only a single dialect, while the others may be represented by two (including Tsou, Saisiyat and Pazeh) or more (including all the other extant languages and all extinct

11 Let it be noted that there are, in some cases, more different reflexes for the same PAN phoneme in the extinct languages than in the other Formosan languages. For instance, Taokas, Babuza and Hoanya each has 3 to 4 different reflexes for PAN *D (see Li 1985 "Formosan reflexes of Proto-Austronesian.") This seems to indicate that each of the extinct languages is represented by different dialects from miscellaneous sources, rather than truly having so many different reflexes for the same phoneme.

12 The relationship between Thao and Bunun is most dubious. They share no phonological innovations. They differ in both morphology and syntax to a great extent. Some of the lexical evidence, therefore, can be taken as Thao borrowing from Bunun.

languages). If a cognate is not retained in one dialect, it may be kept in another. Consequently, when a language has several divergent dialects, we are apt to identify a larger number of cognates for such a language. The results of such a study may be skewed to a certain extent.

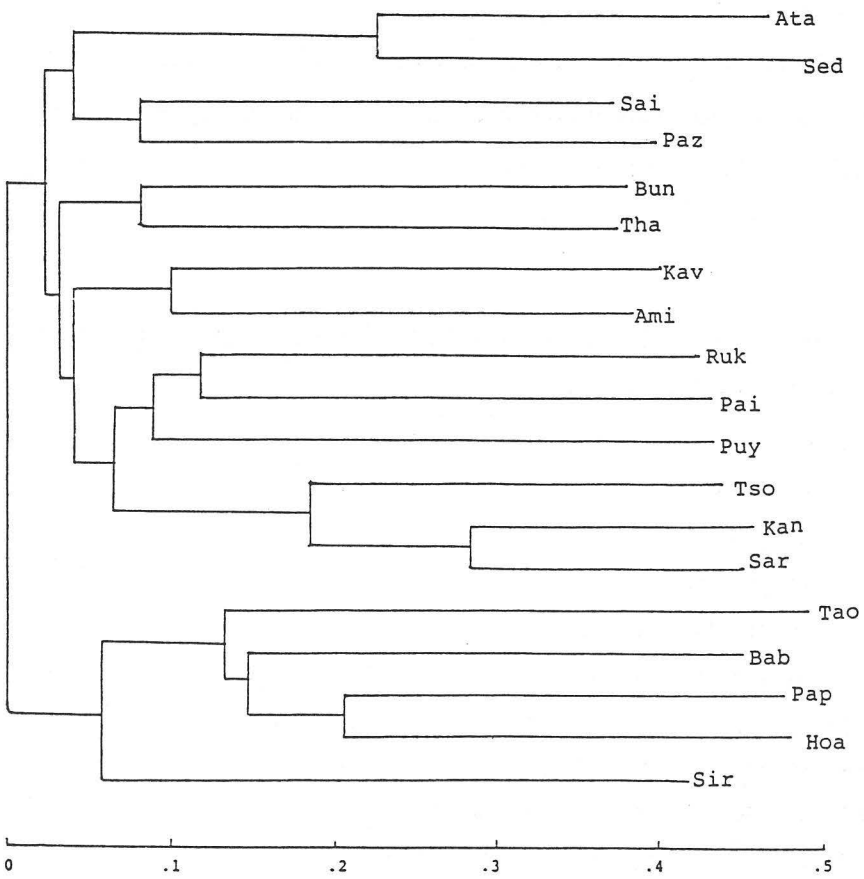
To safeguard against any possible deviations, we have also calculated the number of cognates shared by each pair of languages, in which each language is represented by only a single dialect, as shown in Table 6 below:

Following the same procedure, we can come up with another family tree, as shown in Figure 3.

Table 6. Number of Cognates Shared by Each Pair of Languages Represented by Only a Single Dialect

	Ata	Sed	Tso	Kan	Sar	Puk	Bun	Pai	Puy	Tha	Sai	Paz	Kav	Ami	Tao	Bab	Pap	Hoa	Sir
Ata	293																		
Sed	191	281																	
Tso	75	64	300																
Kan	83	67	217	379															
Sar	81	68	219	299	374														
Ruk	69	64	109	119	129	219													
Bun	82	73	112	116	118	89	211												
Pai	66	61	125	139	143	139	101												
Puy	63	48	99	99	102	98	81	128	220										
Tha	77	70	100	102	101	83	121	96	84	207									
Sai	120	91	114	116	121	97	111	110	94	120	267								
Paz	91	71	87	91	87	72	86	87	78	106	136	222							
Kav	74	60	94	95	92	87	81	92	83	85	99	79	182						
Ami	73	65	98	102	103	91	98	105	97	100	107	87	112	202					
Tao	25	24	20	24	25	19	27	25	24	33	39	40	26	30	85				
Bab	27	28	36	37	40	33	40	43	39	47	43	46	33	38	48	99			
Pap	25	25	29	30	31	27	30	37	31	38	36	42	29	38	36	48	83		
Hoa	20	18	30	34	34	24	33	31	30	37	29	36	30	31	37	49	52	83	
Sir	34	27	52	62	61	47	47	53	50	53	50	38	45	50	27	39	32	35	95

Figure 3. Formosan Classification (Each language is represented only by a single dialect)



Interestingly enough, Figures 2 and 3 are very similar. The main difference between the two lies in the position of Rukai. Figure 2 shows that Rukai groups with the Tsouic languages when each language is represented by one or more dialects, whereas Figure 3 shows that Rukai is most closely related to Paiwan when each language is represented by only a single dialect (Budai, in the case of Rukai). [13] As mentioned above, Rukai is comprised of three divergent groups having six different dialects with considerable lexical,

13 A word of caution: Budai has been surrounded and heavily influenced by Paiwan.

phonological and syntactic differences. It makes a real difference in classification whether we treat Ruaki as a single language or not.

Both Figures 2 and 3 show the close clustering of Atayal and Sediq, Kanakanavu and Saaroa, and to a lesser extent the clustering of Tsou and the two Southern Tsouic languages, and of the four extinct northwestern languages. Higher order subgrouping presents a serious problem, as most pairs of languages, including Saisiyat and Pazeh, Bunun and Thao, Kavalan and Amis, Paiwan and Puyuma or Rukai, share a rather short history of common developments, yet a fairly long history of individual developments, as indicated in our lexical evidence.

In short, the results of this study seem more suggestive than conclusive. These are perhaps the best we can achieve for language classification based on lexical evidence. None of the results have come as a real surprise. They roughly agree with what we have found about Formosan classification based on phonological and syntactic evidence.

REFERENCES

Asai, Erin.

--- 1956. Classification of Formosan languages (in Japanese). *Anthropological Society of Japan, Records of the 10th Meeting*, 62-66.

Bloomfield, Leonard.

--- 1933. *Language*. New York: Henry Holt.

Blust, Robert.

--- 1977. The proto-Austronesian pronouns and Austronesian subgrouping : A preliminary report. *Working Papers in Linguistics* 9.2 : 1-15. Honolulu: University of Hawaii.

Cavalli-Sforza, L. L. and A. W. F. Edwards.

--- 1967. Phylogenetic analysis-models and estimation procedures. *American Journal of Human Genetics* 19 : 237-57.

Clifford, D.H.T. and W. Stephenson.

--- 1975. *An Introduction to Numerical Classification*. New York : Academic Press.

Dahl, Otto Christian.

--- 1981. *Early Phonetic and Phonemic Changes in Austronesian*. Oslo : Institute for Comparative Research in Human Culture.

Dyen, Isidore.

--- 1963. The position of the Malayopolynesian languages of Formosa. *Asian Perspectives* 7.1-2 : 261-271.

--- 1971a. The Austronesian languages and proto-Austronesian. Thomas Sebeok, ed., *Current Trends in Linguistics* 8 : 5-54.

--- 1971b. The Austronesian languages of Formosa. In Thomas Sebeok, ed., *Current Trends in Linguistics* 8 : 168-199.

--- 1987a. Some observations of the Formosan languages. A talk given at the Austronesian Circle, University of Hawaii, April 2, 1987.

--- 1987b. The homomerous method of subclassifying related languages. Paper presented

at the XIV International Congress of Linguists, East Berlin, DDR, 10-15 August 1987.

Ferrell, Raleigh.

--- 1969. *Taiwan Aboriginal Groups : Problems in Cultural and Linguistic Classification.* Taipei : Institute of Ethnology, Academia Sinica monograph No.17.

--- 1972. Verb systems in Formosan languages. In Jacqueline M.D. Thomas and Bernot Lucien, eds., *Langues et Techniques, Nature et Societe, Tome I : Approche Linguistique,* 121-128. Paris : Klincksiek.

--- 1979a. Construction markers and subgrouping of Formosan languages. In Nguyen Dang Liem, ed., *Southeast Asian Linguistic Studies* 3 : 99-211. Canberra : *Pacific Linguistics,* C-45.

--- 1979b. Phonological subgrouping of Formosan languages. In Paz B. Naylor, ed., *Austronesian Studies: Papers from the Second Eastern Conference on Austronesian Languages,* 241-254. Ann Arbor : Michigan Papers on South and Southeast Asia, Center for South and Southeast Asian Studies, University of Michigan, No.15.

Fitch, W.M. and E. Margoliash.

--- 1967. Construction of phylogenetic trees. *Science* 155 : 279-84.

Ho, Dah-an.

--- 1978. A comparative study of Paiwan dialects (in Chinese). *Bulletin of the Institute of History and Philology, Academia Sinica (BIHP)* 49.4 : 565-681.

--- 1983. The position of Rukai in the Formosan languages (in Chinese). *BIHP* 54.1 : 121-168.

Li, Paul Jen-kuei.

--- 1977. The internal relationships of Rukai. *BIHP* 48.1 : 1-92.

--- 1978. Linguistic areal features. *Papers in Honor of Professor Wan-li Chu on His Seventieth Birthday,* 475-89. Taipei : Lian-jing Publishing Co.

--- 1981. Reconstruction of proto-Atayalic phonology. *BIHP* 52.2 : 235-301.

--- 1985. The position of Atayal in the Austronesian family. In Andrew Pawley and Lois Carrington, eds., *Austronesian Linguistics at the 15th Pacific Science Congress,*

257-280. *Pacific Linguistics*, C-88.

--- 1988. A comparative study of Bunun dialects. *BIHP* 59.2 : 479-508.

--- Formosan cognates. Unpublished manuscript, pp.160.

Sneath, P.H.A. and R.R. Sokal.

--- 1973. *Numerical Taxonomy*. London : Freeman.

Starosta, Stanley.

--- 1985. Verbal inflection versus deverbal nominalization in PAN : The evidence from Tsou. In Andrew Pawley and Lois Carrington, eds., *Austronesian Linguistics at the 15th Pacific Science Congress*, 281-312. *Pacific Linguistics*, C-88.

Tsuchida, Shigeru.

--- 1976. *Reconstruction of Proto-Tsouic Phonology*. Tokyo : Study of Languages and Cultures of Asia and Africa, monograph series No.5.

--- 1982. *A Comparative Vocabulary of Austronesian Languages of Sinicized Ethnic Groups in Taiwan, Part I: West Taiwan*. Memoirs of the Faculty of Letters, University of Tokyo, No.7.

--- 1983. Austronesian languages in Taiwan (Formosa). In S.A. Wurm and Shiro Hattori, eds., *Language Atlas of the Pacific Area*. Canberra : The Australian National University.

--- 1985. Kulon: Yet another Austronesian language in Taiwan? *Bulletin of the Institute of Ethnology, Academia Sinica* 60 : 1-59.

Wang, William S-Y.

--- 1987. Representing language relationships. In Henry M. Hoenigswald and Linda F. Wiener, eds., *Biological Metaphor and Cladistic Classification : An Interdisciplinary Perspective*, 243-256. Philadelphia : University of Pennsylvania Press.

--- 1989. The migration of the Chinese people and the settlement of Taiwan. In Li, Kuang-chou et al, eds., *Anthropological Studies of the Taiwan Area : Accomplishments and Prospects*, 15-36. Taipei : Department of Anthropology, National Taiwan University.

台灣南島語言的分類：詞彙證據

（摘要）

李 壬 癸

台灣南島語言的分類問題至今仍然沒有令人滿意的解決方案。本文嘗試以作者歷年來所鑑定的同源詞的多寡來計算各語言之間的親疏關係和彼此的距離。

我們的計量方法是根據遺傳學界所發展出來的運算方式，以電腦計算，然後再以人工換算成距離而繪成樹圖。今日仍然存活的語言共有十四種和已消失的五種語言資料都作爲比較研究資料。前者各種語言各約有一千個詞項，而後者各種語言各的有四百個詞項（所鑑定的同源總數約一千一百個）。因爲材料不平衡，所鑑定的同源詞前者自然遠多於後者，在計量方面也就有必要做一些調整，本文採用了Jaccard的「係數原理」。即使經過調整，五種消失語言的同源詞數量仍然偏低，因此表面看起它們自成一群。

惟獨保存在兩種語言之間的同源詞（exclusively shared cognates）最能顯示它們的親密關係。然而，有好多對（pairs）的語言極少或沒有任何這一類的同源詞。因此，我們不能根據這種資料繪成樹圖。我們仍然要以每一對語言之間所保存的同源總數，來計算它們之間的親疏距離。此外，每種語言是以單一方言作代表，還是綜合各種方言的資料來統計同源詞數字，所得的結果也會略有不同：兩種樹圖就顯示魯凱語的歸屬不同。

較低層次的關係（如泰雅和賽德克，卡語和沙語）顯而易見，不成問題。而較高層次問題就很多了，正如樹圖中所顯示的，各組語言的共同歷史很短暫。

本文所採用的方法和傳統的詞彙統計法（lexicostatistics）有所不同。第一，詞彙統計法每次只比較兩種語言，而在本文所有的語言都一起比較，以標準誤差來檢查錯誤。第二，詞統計法只用一百或二百個基本詞彙，而在本文使用所有的詞彙，因爲資料愈多錯誤愈少。

以詞彙的證據作爲語言分類的依據，和以音韻或句法的證據，所得的結果大致相近，但也有一些出入。以不同的證據和方法所得的結果都可作爲分類的參考。本文嘗試一種新的計量方法，希望提供一種新的分類證據。

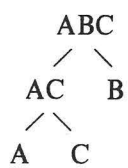
Subgrouping by Lexical Similarity

Stanley Starosta

Dr. Paul Jen-kuei Li's article in this issue, Classification of the Formosan languages: lexical evidence, presents a subgrouping of Formosan aboriginal languages which is constructed by means of a computer program which takes only percentages of shared cognates as input, and produces a genetic classification which groups together those languages which have more cognates in common. Li himself seems rather ambivalent about the value of the new method, but he does clearly assert at several points that the degree of lexical similarity directly reflects the degree of genetic relatedness. I will contend in this note that in fact it may not, and that genetic subgrouping based solely or even mainly on degree of lexical similarity is not valid.

Subgrouping by means of lexical similarity is based on the assumption, supported in Li's article by neither arguments nor evidence, that languages which have a larger number of shared cognates in common must have a longer period of shared history than those which share fewer cognates. Thus if languages A and C share 80% of their cognates with each other, but each shares only 70% of its cognates with language B, then A and C must necessarily be descendents of a single language AC which split off from ABC:

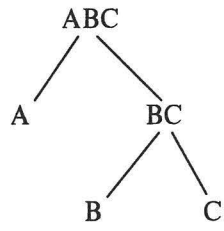
Figure 1.



However, the validity of this method depends on the assumption that the rate of replacement of vocabulary for any language is constant, so that more shared cognates necessarily reflect a longer shared history. If this assumption is incorrect, the lexical similarity

method could give the wrong tree. Suppose for example that in actual historical fact, A separated off first, i.e..

Figure 2.



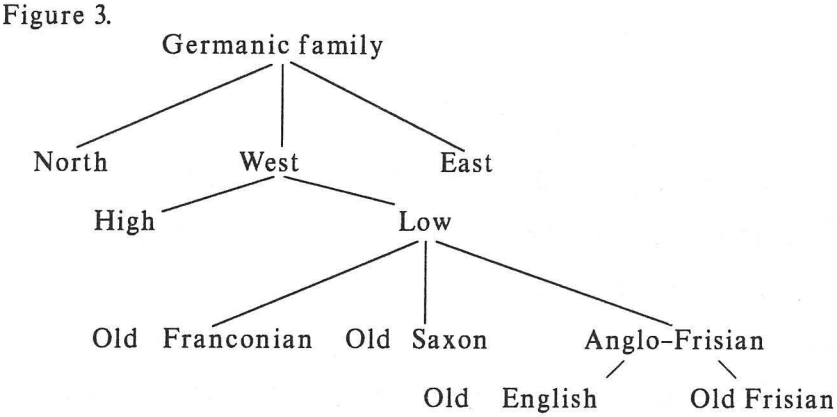
but that C lost its vocabulary more slowly than the others while B lost its vocabulary much more quickly. In that case, A and C would come to have a greater percentage of shard vocabulary with each other than either had to B, and a genetic subcategorization based on lexical similarity alone could produce the tree in Figure 1, a tree which is not in accord with actual historical language splits as depicted in Figure 2. Although the statistical method Li uses compares all the languages at once rather than doing this sort of pairwise comparison, it still has to start off with the percentages of shared vocabulary at a given point in time, and so would have to reach a comparable incorrect result.

So, is the rate of loss of vocabulary constant or isn't it? Even lexicostatistics, which shares similar assumptions with Li's method, has never claimed a constant rate of replacement for anything but 'basic vocabulary', and even that minimal claim is controversial. Li however explicitly abandons this restriction and extends his method to include all vocabulary, not just 'basic vocabulary', despite the fact that, as far as I know, no linguist would claim that the rate of change is constant for all vocabulary. In fact, simple observation should be enough to show that it is not, since political, social, and technological changes may result in the rapid introduction of whole new sets of vocabulary and make other sets suddenly obsolete. That being the case, the rate is not in fact constant, and the lexical similarity method of reconstructing linguistic history is invalid in principle. QED.

Note that one consequence of the lexical similarity method of genetic reconstruction is that subsequent events can change prior history. That is, any event which drastically changes the percentage of shared cognates will result in a redrawing of the genetic

subgrouping tree. Thus a linguist who applied the lexical similarity subgrouping method to the Germanic languages in 1065 A.D. would probably have produced a tree such as Figure 3, which matches the tree drawn by the standard comparative method, showing English and Frisian as sisters.

Figure 3.

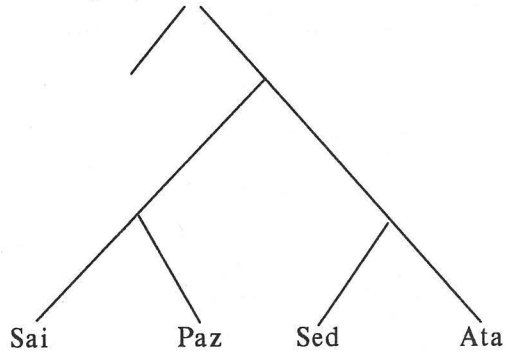


However, a linguist applying the same quantitative methodology today would almost certainly produce a different tree because of the dramatic replacement of native Germanic vocabulary in English by Romance vocabulary beginning with the Norman conquest. In this new tree, English would no longer be an immediate sister of Firsian; rather, the lower cognation percentages would require it to be attached somewhere else higher up in the tree. The lexical similarity methodology would then in effect have made the claim that the Norman Conquest altered prior history. A method of drawing family trees which tells you tonight that Ms. S. Doe is your Sister Sally and then tells you tomorrow morning that she has become your Aunt Sarah does not inspire confidence.

Finally, several practical problems of implementation for this method are clearly pointed out by Li in his article. These are 1) the tendency for geographically adjacent languages to have higher percentages of shared cognates, and 2) the related problem of inter-dialect borrowing and the skewing that results if loans are not clearly distinguished from true cognates. In fact, the Formosan genetic subgrouping tree Li himself presents in this paper appears to exemplify this point. According to Li's computer-drawn tree, Pazeh and

Saisiyat form one subgroup, Atayal and Seediq form a second, and these two subgroups together form a single subgroup:

Figure 4.



According to this subgrouping, Saisiyat and Pazeh are equally close to Atayal and to Seediq. However, the cognate percentages Li presents as the basis for this tree do not bear this out. While Saisiyat and Pazeh do indeed show high cognate percentages with Atayal, their percentages with Seediq are not high at all, and in fact are significantly lower than the percentages obtaining between Saisiyat and Pazeh and a number of languages to the south. This is of course a problem for any methodology depending on percentages of shared cognates. Li's computerized lexical similarity method of drawing family tree does indeed force a resolution of all discrepancies and produce a single unique tree, but this can in no way be considered a solution to the problem. Rather, this approach has only concealed the problem, and in fact it will always necessarily misrepresent some of the actual input data in any tree containing such discrepancies.

So what can be done to resolve this problem satisfactorily? Statistical methods by definition cannot supply the answer. It seems probable that Saisiyat and Pazeh have such high apparent cognate percentages with Atayal simply because they are geographically adjacent to Atayal but not to Seediq, and have borrowed heavily from Atayal. (My own Saisiyat informant for example was bilingual in Atayal, and hardly ever had the occasion to speak Saisiyat in daily life.) To solve this problem, it is necessary to identify all the Atayal borrowings in Saisiyat and Pazeh so that they can be excluded from the lexical similarity

calculations. The only way I can see to go about this is to do more of what Dr. Li has been doing so successfully all these years: looking at more data and applying the classical comparative method to work out the sound correspondences in even more detail, make better hypotheses about proto-forms and shared innovations, and then eliminate from consideration those items of vocabulary which don't fit the scenario. Once this job is finished, though, a reliable genetic subgrouping tree will have been constructed in the process, and the lexical similarity method has nothing new to add.

My conclusion from these considerations is that it is not possible to do genetic subgrouping by feeding percentages into a machine and pulling a switch. It just isn't that easy; there are hard judgement calls to make, decisions that can't be left to a simple mechanical algorithm. At the same time, the method is very attractive, and there should be a place where we can put it to work to draw those nice trees. Li has suggested one such place, that is, in providing an initial approximation of a genetic subgrouping tree, to be revised and improved as more information on shared innovation is worked out. Another application might be to regard the results of the program as simply a quantification of the degree of lexical similarity between two languages, but NOT as a measure of genetic relatedness. It would then not be necessary to distinguish cognates from borrowings at all. Instead, we would simply include all similar-looking words in the lists and output a lexical similarity tree which would allow us to say that, for example, Italian and Spanish are lexically more similar to each other than either is to French (if they are), or that English and Japanese are lexically more similar to each other than either is to Mongolian (if that turns out to be the case).

Reply to Dr. Starosta's Comments

Paul Jen-kuei Li

I have profited from Professor Starosta's comments on an earlier manuscript of this paper. Moreover, I appreciate his remarks on the paper I have presented here. He has raised some interesting questions to which there are no easy answers.

I have applied the standard comparative method to the study of Formosan languages for the past twenty some years with rather fruitful results. I have used it to determine all the cognates in these languages in both my previous and present research (see Section 4 in this paper). However, the standard comparative method has its limitations. One is that it is often not helpful for higher order subgrouping. Another is that it may fail to distinguish between true cognates from loans especially at an early stage. Nevertheless, the new method of subgrouping still must utilize cognates that were identified by the comparative method.

A new method of subgrouping is applied to the study of Formosan languages in this paper. The results are checked against those achieved through the traditional comparative method or simply by inspection. The results obtained by the different approaches often confirm or complement each other, rather than contradict each other. We have not gone so far as to claim that the new method can replace the standard comparative method.

Professor Starosta has correctly pointed out the problem, just as I have indicated in my paper (see Sections 4 and 6), that in some cases (such as Saisiyat, Pazeh and Atayal) I may have failed to completely distinguish true cognates from loans in the input data. Most cognates have been determined on the basis of regular sound correspondences following the procedure of the standard comparative method. It is sometimes difficult to determine whether a certain form in a language is a cognate or not. I have, in fact, weeded out quite a few borrowings between languages. However, if borrowings take place in the forms that

contain the segments (consonants or vowels) which have not undergone sound change, the comparative method cannot determine whether these are loans or true cognates. Naturally more careful and thorough work needs to be done on the historical developments of all these languages. When more reliable data become available, we may come up with a somewhat different subgrouping tree.

Professor Starosta seems to have underestimated the value of Formosan cognates that I have taken pains to identify all these years. I believe these cognates must have a bearing for genetic closeness of these languages. It seems to me the value of true cognates is much higher than that of loans. A tree showing lexical similarity between languages due to loans is of little interest for genetic classification, a case in point being Japanese and Chinese. Japanese had extensive borrowing from Chinese in history, yet they belong to different language families, i.e. Altaic and Sino-Tibetan respectively.

Comments on Subgrouping by Lexical Similarity

William S. Y. Wang

These 4 points occur to me after reading Starosta's comments on lexical similarity (hereafter LS). I would be happy to discuss them with either of you at any time.

(1)The comparative method is rife with unsolved problems. These have been discussed by many authors. It is by no means a panacea for detecting historical relations. Perhaps the most famous discussion is by Bloomfield (1933:316), in which he noted the difficulty of drawing trees for the Indo-European languages. As students of historical linguistics, it is our task to explore ways to refine, complement, supplement, or replace the comparative method. Such explorations are necessary for the growth of historical linguistics. LS is at an early stage of development in linguistics, though it has been applied successfully in biological phylogeny for many more years.

(2)Let us assume with Starosta's Figure 2 that the true history is (A (BC)), i.e., A was the first language to split off. Let us also assume that A replaces its vocabulary slowly while C does it more quickly. The difference in rates of vocabulary replacement would not in itself result in Starosta's Figure 1, i.e. ((AB) C). Such a result would only come about if there are selective forces in C that cause the replacement of a disproportionately large number of cognates which C shares with B. This hypothetical scenario only carries force if it can be demonstrated that such selective forces are often at work to a significant extent. To my knowledge, no such demonstration is available.

(3) No-one can answer the question as yet of "at how constant a rate do languages replace their vocabulary?" Hopefully, more studies of LS can shed some light on this question. Unlike glottochronology, LS does not assume a constant rate. In fact, all LS results have shown that rates vary from language to language. These variations are

explicitly indicated by the length variations in the branches of LS trees. Furthermore, it should be noted that the branch lengths are relative to each other, rather than absolute in time. Its temporal value needs to be interpreted by calibration against external information. This is a critical difference between glottochronology and LS. Another difference is that LS does not commit to a "basic vocabulary", the delimitation of which is extremely difficult and sometimes arbitrary. Yet another difference is that by proceeding pairwise, glottochronology loses much of the information very useful for larger groupings. One must not confuse LS with glottochronology.

(4) The point is well taken that massive borrowings can complicate the historical picture, for LS as well as for any other method, including the comparative method. However, in the present case, the vocabulary has been carefully scrutinized and sifted for true cognacy by a recognized authority in the field, i. e., Paul Li. Borrowing across languages is not restricted to vocabulary, of course. There is some indication that LS perhaps can be refined eventually to yield information on the relative amounts of vertical versus horizontal transmission.