7

The position of the Malayopolynesian Languages of Formosa

ISIDORE DYEN

Yale University

It is of some interest to know how many different families of Malayopolynesian languages are to be found in Formosa and to determine their relationship with other Malayopolynesian families.[1] Although the material now available does not permit a final decision, there appears to be some indication of a trend.

My article 'The Lexicostatistical Classification of the Malayopolynesian Languages' (Dyen 1962) shows that, a number of the Malayopolynesian Languages of

[1] This article is based on work supported at different stages by grants by the Tri-Institutional Pacific Program and by the National Science Foundation. The following is a list of abbreviations of languages and bibliography used in this article:

Am_1 — Ami LL, E. P. Torgesen.
Am_2 — Ami LL, Chen-Li.
Am_a — Ami, all dialects, OA.
AmB — Ami, Baran dialect, OA.
AmK — Ami, Kibi dialect, OA.
AmTa — Ami, Taparon dialect, OA.
AmTo — Ami, Tooran dialect, OA.
At_1 — Atayal LL, Clare McGill.
At_2 — Atayal LL, CL.
AtTh — Atayal (Taihyoo dialect), OA.
AtTk — Atayal (Takonan dialect), OA.
Bu_1 — Bunun LL, Fu Wen-Chih.
Bu_2 — Bunun LL, CL.
BuN — Bunun, northern dialect, OA.
BuC — Bunun, central dialect, OA.
BuS — Bunun, southern dialect, OA.
CL — set of lexicostatistical lists submitted by Chen Chi-lu on the basis of material of F. K. Li.
Ilk — Ilocano.
K — Kanabu or Kankanabu language, OA.
Kv — Kavalan LL, CL.
LL — lexicostatistical list.
Ml — Malay.
OA — N. Ogawa and E. Asai (1935).
Pa_1 — Paiwan LL, John Whitehorn.
Pa_a — Paiwan, all dialects, OA.
PaDm — Paiwan, Daima dialect, OA.
PaDt — Paiwan, Daityooman dialect, OA.
PaKp — Paiwan, Kapiayn dialect, OA.
PaKt — Paiwan, Katirai dialect, OA.
PaKu — Paiwan, Kunanau dialect, OA.
PaLi — Paiwan, Likiliki dialect, OA.
PaLP — Paiwan, Lower Paiwan dialect, OA.
PaN — Paiwan, Nai dialect, OA.
PaNb — Paiwan, Naibun dialect, OA.
PaT — Paiwan, Tokubun dialect, OA.
Pu_2 — Puyuma LL, CL.
Pu_a — Puyuma, both dialects, OA.
PuT — Puyuma, Tippon dialect, OA.
PuP — Puyuma, Pinan dialect, OA.
Pz — Pazeh LL, CL.
R_2 — Rukai LL, CL.
R_a — Rukai, all dialects, OA.
RMg — Rukai, Maga dialect, OA.
RMn — Rukai, Mantauran dialect, OA.
RTn — Rukai, Tainan dialect, OA.
RTo — Rukai, Tona dialect, OA.
RTr — Rukai, Taramakau dialect, OA.
Se_1 — Seedik LL (Taroko dialect), Ralph Covell.
Se_2 — Seedik LL, CL.
SeK — Seedik, Kiri dialect, OA.
SeT — Seedik, Taroko dialect, OA.
Sr — Saaroa, OA.
Ss — Saisiyat, OA.
T_a — Tsou, both dialects, OA.
TA — Tsou, Arizan dialect, OA.
Tg — Tagalog.
Th — Thao LL, CL.
TR — Tsou, Rufuto dialect, OA.
TT — Tsou as reported in T. H. Tung,(n.d.)

Words are often cited in a different form from that of the source; the substitutions, where not one-for-one, are in accordance with my phonemic interpretation of OA.

Formosa have a critical percentage of less than 25%; they are: Thao (24·5%), Bunun (24·5%), Paiwan (24·5%), Ami (24·5%), and the Atayalic Subfamily ([there called Group] 13·6%). In the case of an individual language, the critical percentage used here is the highest percentage with any language; for a group, it is the highest percentage of any member with a non-member language. The Atayalic subfamily has as its members Atayal and Seedik whose lists exhibited 35·2% cognation.

To these languages can now probably be added Tsou. The highest percentage of Tsou with a Formosan language is 16·7% which it holds with Bunun and with Paiwan.

The lexicostatistical evidence given in my article (Dyen 1962: 38–46) indicates a great preponderance of highly diverse languages in the Melanesia-New Guinea area and thus implies that this area is the Malayopolynesian (or Austronesian) homeland; that from this area seven have a critical percentage 15·0% or lower and only one elsewhere; and furthermore from this area twenty-five (and seven from West Irian) have a critical percentage of 20·0% or less and only two elsewhere than in this area or West Irian.

Now the lowest critical percentage of any language listed in the article is 11·1% for Hapa of Morobe, NE New Guinea. Thus the percentages for the Atayal family and for Tsou are interesting because they suggest the possibility that Formosa rather than Melanesia-New Guinea may be the locus of the most diverse Malayopolynesian languages. This view could be defended on the grounds that the lexicostatistical percentages below 20·0% (and perhaps even those below some larger figure) are not sufficiently different to be taken at face value. It is not at all inconceivable that all of the Melanesia-New Guinea percentages are lower than they should be, if their proper relation were indicated. In that case the area of the Formosan languages might be proposed as a possible homeland because of the low scores of the Atayalic Subfamily and Tsou.

The argument presented here is that an examination of the vocabularies of the Formosan languages as they appear in the available material does not seem to confirm this hypothesis. Rather the indications from the vocabularies are, thus far, that the Formosan languages more probably constitute a single family.

Ogawa and Asai (OA) (1935: 3f) list the following languages of Formosa:

1. Atayal
2. Seedik
3. Saisiyat
4. Bunun
5. Tsou
6. Kanabu (=Kankanabu)
7. Saaroa (=Sisyaban)
8. Rukai
9. Paiwan
10. Puyuma
11. Ami
12. Kavalan
13. Pazeh
14. Thao (=Sao)
15. Ketagalan
16. Taokas
17. Papora
18. Babuza (=Favorlang)
19. Hoanya
20. Siraya

Of these numbers 15–20 belong to dead or nearly dead languages. In 1935 they were not spoken as native languages or were only remembered by old people.

Of the remainder the first eleven are reported on comparatively fully by OA. They also report on Yami, the language of Botel Tobago, but this is evidently closely related to Ivatan, the language of the Batan Islands. The latter seems clearly to be a member of the Philippine group of languages. For this reason we shall not discuss the Yami language.

TABLE I

At_1	59·0													
At_2	60·0	82·9												
Se_1	32·7	37·7	32·7											
TT	07·4	09·0	07·8	10·9										
Th	06·6	08·3	07·4	13·5	16·3									
Bu_1	07·8	07·6	06·8	12·2	15·2	28·8								
Bu_2	08·3	08·3	08·0	11·7	16·7	29·2	65·7							
Pz	10·3	08·9	08·6	14·6	14·3	23·0	20·0	16·1						
Am_1	09·7	08·9	08·7	12·9	15·7	25·9	25·4	27·0	19·9					
Am_2	08·7	08·9	08·4	13·2	14·6	25·5	24·0	23·5	20·7	75·1				
Kv	06·2	06·9	06·1	10·0	13·9	19·0	18·9	19·0	18·5	24·7	24·1			
Pu_2	08·7	07·5	07·8	12·9	14·7	22·5	21·2	21·2	22·0	29·8	28·3	19·9		
R_2	07·9	06·8	06·0	09·9	13·1	16·7	17·4	15·4	14·8	20·0	20·1	15·4	25·6	
Pa_1	09·4	09·3	08·0	15·8	16·7	24·4	23·7	24·5	23·5	27·3	27·4	20·3	28·7	23·8
	Se_2	At_1	At_2	Se_1	TT	Th	Bu_1	Bu_2	Pz	Am_1	Am_2	Kv	Pu_2	R_2

Finally OA say of Kanabu (6) and Saaroa (7) that they should be regarded as dialects of Tsou (Ogawa and Asai 1935: 3f). They say the same of the relation between Seedik and Atayal *(idem)*, but it appears more likely that their relation is that of closely related languages than of dialects of the same language. Perhaps the relation indicated between Tsou and the other languages is likewise that between relatively closely related languages.

Swadesh lists of the following Formosan languages are available to me: Atayal (At_1, At_2), Seedik (Se_1, Se_2), Bunun (Bu_1, Bu_2), Rukai (R_2), Paiwan (Pa_1), Puyuma (Pu_2), Ami (Am_1, Am_2), Kavalan (K), Pazeh (Pz), and Thao (Th). The results of a new lexicostatistical comparison (by hand) is given in Table 1.[2] These results suggest that at the highest level there are three language groups in this comparison: F_1 containing Atayal and Seekik, F_2 containing Tsou, and F_3 containing the remaining languages.

The two lists of Seedik score significantly differently from each other. Se_2 is very much like an Atayal language; it scores 59·0% with one and 60·0% with the other Atayal list. On the other hand Se_1 scores 37·7% and 32·7% respectively with the same lists and 32·7% with Se_2. Furthermore Se_1, as might be expected from the fact that it is from the Taroko dialect, most commonly agrees with the Seedik of Taroko village (SeT) as recorded in OA. On the other hand Se_2 quite frequently

[2] The lists available to me at the time of the machine calculation represented in my article in Language 38 are: At_1, Se_1, Th, Bu_1, Am_1, Pa_1. The following table presents the old percentages as calculated by machine and the new ones obtained by hand:

agrees in form with the Atayal of Takonan as recorded in OA. I infer that Se_2 has been misnamed.

Group F_3 is determined by the following percentages:

Th — Bu_2	29·2	Pu_2 — Pa_1	28·7
Bu_2 — Am_1	27·0	Pu_2 — R_2	25·6
Am_1 — Kv	24·7	Pu_2 — Pz	23·5
Am_1 — Pa_1	29·8		

On the principle of subgroup homogeneity Atayal would be expected to show approximately the same percentages as Seedik (=Se_1). However, with members of F_3, Se_1 averages 5·1 percentage points (pp) higher than At_2 and 4·5 pp higher than At_1. Table 2 presents the differences between the percentages of Se_1 and At_2 with a given language (L_i) and the differences between the percentages of Se_1 and At_1 with the same language:

TABLE 2

	$(Se_1–L_i)$—$(At_2–L_i)$	$(Se_1–L_i)$—$(At_1–L_i)$
Th	6·1	5·2
Bu_1	5·4	4·6
Bu_2	3·7	3·4
Pz	6·0	5·7
Am_1	4·2	4·0
Am_2	4·8	4·3
Kv	3·9	3·1
Pu	5·1	5·4
R	3·9	3·1
Pa	7·8	6·5
	51·0	45·3
Average	5·1	4·5

This creates a suspicion that perhaps Atayal exhibits distorted percentages.

Qualitative evidence for the existence of a Proto-Formosan can take the form of likely common Formosan innovations: that is, sets of likely cognates which are found in the putative Formosan branches and have no extra-Formosan cognates among Malayopolynesian languages. The following sets of cognates satisfy these requirements; they exhibit at least one member from F_1 and one from F_3:

	Old	*New*		*Old*	*New*
At_1 — Se_1	35·2	37·7	Se_1 — Pa_1	13·6	15·8
At_1 — Th	08·1	08·3	Th — Bu_1	24·5	28·8
At_1 — Bu_1	08·5	07·6	Th — Am_1	19·8	25·9
At_1 — Am_1	10·4	08·9	Th — Pa_1	20·9	24·4
At_1 — Pa_1	10·1	09·3	Bu_1 — Am_1	23·0	25·4
Se_1 — Th	12·5	13·5	Bu_1 — Pa_1	20·3	23·7
Se_1 — Bu_1	11·5	12·2	Am_1 — Pa_1	24·5	27·3
Se_1 — Am_1	12·5	12·9			

The difference generally favours the later hand determination: this suggests that, if a difference in attention is not reflected, the linguist has perhaps gained some experience. Nevertheless it should be noted that the new figures only tend to emphasize a trend observable in the machine figures.

1.1. At_2 pali?, AtTh pali, Bu_2 pani?, BuC, BuS pane 'feather'.

1.2. AtTk raho?, TRn, RTr, R_2 maḍau 'big'.

1.3. At_2 ramu, AtTh ramo, AtTk ramu-juX, Pa_1 jamoq, PaT ḍamo?, PaP damok, Pu_2 ?ada·moq, Pz ?damo? 'blood'.

1.4. At_1 yuluŋ, At_2 juluŋ, AtTh, AtTk juluŋ, Se_1 roloŋ, SeT ruluŋ, RMg krororŋo, Sr ḷoḷoŋa 'cloud'.

1.5. Se_1 bělěbil, Pa_1 venilevil, R_2 bilbil 'pull'.

1.6. At_1 měqwas, pěqwas, At_2 məquas, AtTh məqowas, AtTk ma?owas, SeK uwes, SeT ujas, Se_1 oyas (mooyas), Th maqa-qú·jiš, Bu_1 ka-xúdas, BuS xođas 'sing'.

1.7. At_1 hěma-li, At_2 ḥəma-di, AtTh hama-li, AtTk hama-?oi, SeK he:ma, SeT xamma/hmma, Se_1 hěma, Pa_1 sema, PaLP sma, PuP sīma, Am_a sīma 'tongue'.

1.8. At_1, At_2, AtTh, SeK, SeT, Se_1 kawas, Th ká·waš, Pz qawas 'year'.

1.9. AtTh hera, AtTk so:-hesa, RTr ko-eja, RMn iđa, RTo haku-si?a, Am_a ina-tsiɹa 'yesterday'.

1.10. AtTh mipusal, AtTk maposan, SeT mappusal, RTr maposal, R_2 mausaḷ, Bu_a mapuşan, K mapusanu, Sr mapuwaḷu, Th mapušađ 'twenty'.

1.11. AtTh juŋai, AtTK zuŋai, SeK loŋŋai, SeT loŋŋaĭ, PuT uŋai, AmTo o?ŋai 'monkey'.

1.12. AtTh qaom, AtTk aɣom, SeT aḷom, PaT ?a:m, PuT harum, PuP arum, Am_a aɹum, BuC qalom, BuS xaḷom, Sr ?arumu 'ant eater'.

1.13. SeK, SeT, RTn, Pu_a walo, BuN wanno, BuC, BuS vanno, K anu, Sr ?alo?o 'bee'.

1.14. AtTk rotok, Ss ɣotok, PaT ḷotok, RTr ḷutuk^u, T_a jutuka, 'rabbit'.

1.15. AtTh ḅətakan, AtTK ḅatakan 'big bamboo', SeK butakkan 'bamboo', BuC, BuS ḅatakan 'big bamboo', TR pu?tsoknu, TA putsoknu 'bamboo'.

1.16. AtTh hajuŋ, SeK haroŋ, SeT haḷoŋ, PaT talīŋ, RTn ḥaḷiŋ^i, Am tsaḷiŋ, BuS saaŋ, TA seoŋu, K aḷuŋu, Sr (h)aḷuŋu 'pine'.

1.17. AtTh, AtTK kai, SeK kairi, SeT ka:ḷi, PaT kai, K ka:ri, Sr kari 'language'.

1.18. SeK blebl, SeT bulbul, Pa_a vulvul, RTr bulbul, PuT vulvul, PuP bulbul, Bu_a bun-bun, K ta-ḅunuḅunu, Sr ta-ḅuḷuḅuḷu 'banana'.

1.19. SeK wassaŭ, PaT asao, RTr vasau 'leaf'.

1.20. SeK, SeT buŋa, PuP boŋa RTa, RTo voŋa 'sweet-potato'.

1.21. Se_1 sěpog (sěmpog), Th šmu·pił, Bu_2 mašipul, Pa_1 semopo 'count'.

1.22. SeK bakki, SeT bakke, AmB faki, AmTo vaki 'grandfather'.

1.23. SeK idas, SeT i:das, PaT ?ilas, PaLP qilas, PuT ḥelas, Pz ?ilas 'moon'.

1.24. SeT daleḥ, Se_1 dalih, RMg me-dali, Sr ma-sa:ḷi 'near'.

1.25. SeK ukka, BuC, BuS ʊkka, TR uk?a, TA uk(a)?a, Sr uka?a 'not exist'.

1.26. Se_1 hoda, Pa_1 sola, Th ?úlda?, Am_1 sorla, Am_2 suh da?, Kv suɣona?, Pu_2 ?orla? 'snow'.

1.27. SeK broa, SeT bl̥uwa, BuS bil̥va 'thunder'.

1.28. SeK rebu, T_a sifu, K ibu, Sr i:bu 'urine'.

1.29. At_1 royiq, At_2 rujɛq, AtTh raojeq, AtTK raoji, Se_1 daoriq, SeK doṛeq, SeT daŭl̥eq, Pz ʔdaurik 'eye'.

1.30. At_1, At_2 timu, AtTh tsimo, AtTK tumu-jux, Se_1 timo, SeK tsim̥o, SeT tsimo, RTn_1, RTr_2 timo, RMg timus, RTo timoso, RMn timo 'salt'.

1.31. Se_1 balay, Th muʔbalaʔbá·laj 'right (correct)'.

1.32. At_1 siyik, AtTh səjik, Th ríši? 'liver'.

1.33. At_1 giqas, AtTh geqas, Pz hijaš 'new'.

1.34. At_2 məstəmaq, Pu_2 maṭəməʔ 'rotten'.

1.35. AtTh utas, SeT uttas, PuP ʔtas 'penis'.

1.36. At_1 ruma, Se_1 doma, Pa_1 dzoma 'some'.

1.37. Se_1 sinaq (sĕminaq), Pa_a semenaq 'wash', K ma-tsina 'wash, wash body', TA mamtsi:no, Sr ma:sinu 'wash body'.

Among those listed here a notable number have the following basic meanings: 1.1 'feather', 1.2 'big', 1.3 'blood', 1.4 'cloud', 1.5 'pull', 1.6 'sing', 1.7 'tongue', 1.8 'year', 1.19 'leaf', 1.21 'count', 1.24 'near', 1.29 'eye', 1.30 'salt', 1.31 'right (correct)', 1.32 'liver', 1.33 'new', 1.34 'rotten', 1.36 'some', 1.37 'wash'. This is not to say that some of the other meanings are not just as valuable for this purpose, or that those listed here are necessarily all more valuable than the others; no matter how one regards the individual members of the whole collection, the fact that there is so large a number of probable cognates in meanings in which borrowings as a rule are infrequent makes it probable that we are dealing with a subcollection of cognates that connect the Atayalic family with other Formosan languages.

The implication is then that the percentages of not only Atayal, but also of Seedik are lower than they should be. We conclude that the closest relative of the Atayalic Subfamily is to be found among the other Formosan languages, perhaps in F_3, and this subfamily is not likely to prove to be equally related with F_3 and extra-Formosan Malayopolynesian subfamilies.

The highest percentage of Tsou is 16·7% with Bu_2 and with Paiwan. Tsou on superficial examination appears to be quite aberrant. Thus it would be interesting to know whether there are cognate sets limited to Formosan languages which include a Tsou member.

The following list which is restricted to Tsou, Kanabu and Saaroa, shows that the last two form a group, but the evidence is not lexicostatistical. Whether they form a group or not, the following comparisons suggest a connection between the three languages:

2.1. T_a hitsu, Sr il̥itsʊ 'spirit'.

2.2. TR tşonni, TA tsoni 'one', K tsani 'one'.

2.3. TR humulu, TA humuju, K nimuruʔu 'blood'.

2.4. TR povoʔu, TA povʔū, Sr puwakū 'wing'.

2.5. TA tsūmū, Sr ta:somu 'thorn'.

2.6. T_a tsojowa, Sr sasarowana 'earth'.
2.7. TR, TA majumu, Sr marumu 'bitter'.
2.8. T_a sojumu, Sr masarumu 'cold'.
2.9. TR mutsu, TA emutsu, K ḷamutsu, Sr ḷamutsu 'hand'.
2.10. TR ŋuʔxoŭ, TA ŋuʔhoŭ, ŋhoŭ, K ŋkaŭ 'monkey'.

The following are cognate sets with a Tsou member and one from F_3 and which in addition are restricted to Formosan languages as far as is known:

3.1. T_a tseopuŋu, K tsarapuŋu, Sr sarapoŋu, PaT taḷopoŋ, RTr taḷopoŋ 'hat'.
3.2. T_a fuŋuu, K naƀuŋu, Sr ƀoŋoʔo, AmB fuŋoh, BuN ƀoŋŋo 'head'.
3.3. T_a t'uhu, RTo tūtūnū 'liver'.
3.4. TR tojotsū, TA teotsu, RTn da:tso, PuP ḍarato 'hair-louse'.
3.5. T_a kutʔi, PaT kuti, PuP koti, BuC kutte 'vagina'.
3.6. T_a pasuŋi, K ma'asuŋu, Sr maauŋu, PaT vinaʔsiŋ, RTn wa-basiŋ, PuT paheŋ, AmTa vaʔsiŋ, BuC qasƀiŋ 'sneeze'.
3.7. TR pasunaĭnu, TA pasnaĭnu, Pa_a siminai, RTr wa-sinai, Pu_a siminai 'sing'.
3.8. T_a jeŋhova, Sr maraŋiḷu, AmB ḷaŋḷau 'blue'.
3.9. TR ŋutsu, TA ŋutsu, K taŋutsa, Bu ŋuttuts 'nose'.
3.10. T_a futsuju, PaDm vitsinun, PuT vutinon 'egg'.
3.11. TR timujo, TA timejo, K atimoə, Sr ʔatimuḷa, PuP timuḷa, Am atimuḷa 'flea'.
3.12. T_a mimo, K mima, Sr mijama, SkK, SeT mimaḥ 'drink'.
3.13. TR tso:bixi, TA tsoʔvihi, BuC ḍaqvisan 'far'.
3.14. TA tse:i, Am_a ni-teʔeh 'dream'.
3.15. TT talu, K tumat-im-ana, Sr tuma t-im-aḷa, Bu_a tan'a, Th t-unm-áđaʔ, Pz t-um-araʔ 'hear'.
3.16. TR sofu, TA sofu, PuT saḥov, RTn aob 'roof'.
3.17. TT hisi, K anisi, Sr aḷi, Pa_a alis, RTn walisi, Pu_a wali, AmB waḷis 'tooth'.

To these are to be added 1.14 'rabbit', 1.15 'bamboo', 1.16 'pine', 1.25 'not exist', and 1.37 'wash' from earlier comparisons. There are thus 22 instances to be counted here.

In the preceding two sets of likely cognates the following standard basic meanings are represented: 1.37 'wash', 2.2 'one', 2.3 'blood', 2.4 'wing', 2.6 'earth', 2.8 'cold', 2.9 'hand', 3.2 'head', 3.3 'liver', 3.7 'sing', 3.9 'nose', 3.10 'egg', 3.12 'drink', 3.13 'far', 3.15 'hear', 3.17 'tooth'.

Let us call the cognate-sets with initial numbers 2 and 3 Tsou Formosa-limited cognate sets. It would be risky, to say the least, to ignore these comparisons as perhaps containing material which suggests that Tsou is more closely related with F_3 than with any extra-Formosan language or subfamily. Furthermore Tsou's lexicostatistical percentage is not so low that we are forced to believe that Tsou is to be set off against all the other Malayopolynesian languages.

Not only do the sets of likely cognates given above point to Tsou's membership in a Proto-Formosan, but this same relation is indicated by some non-Formosa-limited cognate sets. The collection is as follows:

4.1. T_a tsufuʔo, Sr tsibuka 'belly' (from *bituka).

4.2. TR fuʔusū, TA fuʔusu, K bukusu, Sr bukuu, AmB fukis, Pz bekụs, Kv boqas, Ss bukis 'hair' (from *buhuk).

4.3. T_a huʔo, Sr ʔaluku 'nail', AmB kanoʔos 'claw'.

4.4. T_a, AtTh iso, SeK issu, BuC soo, Am_1 kiso 'thou'.

Comparisons 4.1 and 4.2 involve metatheses which are limited to the languages cited as far as I know. Compare Tg. bitu:ka, Toba-Batak butuka 'intestines' with 4.1 and Tg. buhok '(head) hair', Javanese wòq 'beard' with 4.2.

Comparison 4.3 is perhaps in some way connected with the widespread set of cognates which indicate both an initial and medial *k: *cf.* Tg. kuko, Ml. kuku 'nail' among others. If the matching phonemes of the Tsou and agreeing words do not differ because of the shared dissimilation, then they indicate the presence of different proto-phoneme instead of *k in one of the two positions (probably the intervocalic position) or that etymon is independent of that of the Tagalog and Malay words.

The words in comparison 4.4 are undoubtedly cognate with Tg. qiyo 'your, you (oblique)'. The use of this form as subject suggests a common innovation among the languages listed.

The Tsou vocabulary appears to point to the fact that its closest relative is probably to be found among the Formosan languages, say those of F_3. This would suggest that Tsou's percentages are lower than they should be. Taken together with the evidence that the closest relative of the Atayalic Subfamily is probably a Formosan language, our evidence now indicates a single Formosan subfamily of Malayopolynesian.

In the course of this investigation the number of comparisons which appeared seemed to agree in pointing toward a close relationship between the Formosan languages as a whole and languages of the Philippines—though no attempt has been made to examine the counter possibilities. It is conceivable that a more careful investigation will reveal just as numerous and cogent etymologies which point to an equivalently close relation with the languages south of the Philippines. However, it is interesting to note that at least this putative collection did not impress itself in the same measure as the connection with Philippine languages. The following are the comparisons referred to above:

5.1. Bu_1, Bu_2 ámin, Am_1 ʔmin, Ilk. qa:min 'all'.

5.2. Pz_2 lamik 'cold', Ilk. lamʔek 'cold'.

5.3. PuT đapal, PuP dapal, Sr l̥apal̥u, Kv z̥apan 'foot', Ilk. dapan 'sole'.

5.4. Th punuq, Pz punoq 'head', HlBs. pu:noq 'head'.

5.5. AtTh səmaqeis, AtTk sumʔis, SeK smaissi, SeT suma-isi, PaT tsimaʔis, RTr wa-tsaisi, AmB mitaʔais, BuC mataqqais, TR t'me'esi, K tumata-isi, Th šmaqis, Pz sasais, Kv təmaʔis, Tg. tahiq 'sew'.

5.6. RTn tsinik^{i}, Tg. tinik 'thorn'.

5.7. Am_a kasui, Pz kahuj, Pu kaui, Tg. ka:hoy 'tree' (cited here only because of the metathesis. Words like Ml kayu 'wood' are almost universal).

5.8. AmB rimakat, Pa_1 jemacat^{s} HlBs. lumakat 'walk'.

5.9. At_1 qemalup, qalup, AtTh qəmalup, Ss 'malup, Pa_1 qemalop, PaT 'umalup, RTr w-alupu, PuT ḥimalup, AmB mi?alup, AmK mi-ḥalup, BuC qonop, BuS xanop 'hunt', Manmanua manganop 'hunt' (with cognates in many other Philippines languages).

The cognates in the last set are only provisionally regarded as limited to the Philippines and Formosa. Any one of them may yet turn out to have cognates outside of these two areas. For example, if a person were not (as for some time I was not) aware of the connection with Ngadju-Dayak eñau 'wash', he would probably also cite here: SeK s-im-inau, RTr wa-sinau 'wash (clothes)', BuC ma-sinav, PuP mu-lisao [with metathesis] 'wash (utensils)', Ilk. ?innaw 'wash (dishes)'. We should of course keep in mind that if—as is not unlikely, but is not taken up as an objective of research—the immediate relation of the Formosan languages as a group is with Hesperonesian (Dyen 1962: 44) and not with the Philippine languages, this relation would explain the limited distribution of the word for 'wash' and presumably some collection of others. Under this hypothesis the highly restricted distribution of the following set of cognates would be explained: SeK nu-noḥ, TR nunu?u, TA nun?u 'breast', Merina nunu 'nipple'. The immediate relationship of the Formosan languages with the Hesperonesian languages would just as effectively reduce the probability of a Formosan homeland of the Malayo-polynesian languages.

The vocabularies of a number of the Formosan languages, which were investigated, appear to contra-indicate that Formosa is the likely homeland of the Malayo-polynesian languages. The argument is essentially that the closest relative of the most diverse languages in Formosa, according to lexicostatistical classification, appears nevertheless in each case to be another Formosan language or subfamily, when the vocabularies of the same languages in Formosa are examined for cognate sets within the same limited range.

Of course, cognate sets of the same limited range are not proof positive of subrelationship even when the number is reasonably large; for such cognate sets can be produced by losses in the surrounding areas and can be simulated by the effects of borrowing. It can however be stipulated that as the number of cognate sets of the same limited range increases, the point of proof positive is approached. The argument from cognate sets of the same limited range is dependent on the probability that the collection of cognate sets contains exclusively shared innovations. Furthermore it should be remembered that not all of the Formosan languages have been subjected to investigation here.

To resort to other phenomena to obtain an evaluation of a lexicostatistical implication may be considered by some to be tantamount to regarding the lexicostatistical evidence as invalid. This is of course not true. However, there are two considerations which prevent us from taking a proportion (or percentage) to have

absolute value. These are *(a)* the fact that a proportion represents a distribution, and *(b)* that a proportion may in a particular case reflect the effect of other factors than are normally present.

Each proportion is the midpoint of a distribution of values and is the most likely value of the distribution. However, if one wishes to consider all of the possible values, which have one chance in twenty or better of being represented by the proportion, this can be done by taking all the values which lie within one and two-thirds[3] deviations of the proportion itself. One standard deviation for Se_1—Pa_1 15·8% based on 165 determinate instances (26 positive instances) is 2·8 pp;[4] one and two-thirds standard deviations are 4·16 pp. Thus there is one chance in twenty (or five chances in 100) that the proportion 15·8% represents a value as high as 20·4%. [There are of course fewer chances than one in twenty that the observed proportion represents some value greater than 20·4%.] Although the same statement could be made for values lying between 11·2% and 15·8% (i.e. within one and two-thirds standard deviations less than 15·8%), the qualitative evidence adduced here do indicate, we think, that these lower values now have a lower probability than otherwise would appear, whereas a value among the higher values has a greater probability of being the true value. What has been said here about Seedik's percentages can likewise be said *mutatis mutandis* about Tsou's percentages.

There is a further possibility that a particular proportion reflects factors other than the normal one of random morpheme decay. We have good reason to believe (Dyen 1963: 60–66) that there is an abnormal (i.e. not present in every case) factor or abnormal factors, among them not unlikely word-taboo, which produce deflated proportions: that is, percentages which are lower than their normal value. We have already cited evidence which indicates that Atayal's percentages are deflated. It is not at all impossible that Seedik's and Tsou's are likewise deflated. If Seedik's percentages are deflated, the implication is that Atayal's are even more deflated than has thus far been indicated.

That Seedik's and Atayal's percentages are possibly deflated does not affect the implication of the low percentages in the Melanesian-New Guinea area. The argument there depends on the large number of languages exhibiting low critical percentages. It is difficult to believe that all or even a significant proportion of these percentages are deflated.

It therefore seems fair to say that one method of checking the lexicostatistical comparison of the Formosan languages gives evidence to suggest that the surprisingly low critical percentages of Formosan languages should reasonably be expected to be corrected upwards to agree with the hypothesis of a homeland in the Melanesia-New Guinea area. Of course it is understood that we now expect that the same method as applied to languages in the Melanesia-New Guinea area will not suffice to produce evidence for a subfamily including all of the languages of that area.

[3] R. A. Fisher and F. Yates (*Statistical Tables*, New York, 1957, p. 42) give 1·64 standard deviations instead of the 'one and two-thirds standard deviations' used here.

[4] From the formula $\sqrt{npq}$ for the standard deviation of the binomial distribution; this leads to $\sqrt{165 \cdot 15{\cdot}8 \cdot 84{\cdot}2} \div 165$ for one standard deviation for the proportion 15·8%.

References

DYEN, ISIDORE
1962 The lexicostatistical classification of the Malayopolynesian languages, *Language* **38**: 34–46.
1963 Lexicostatistically determined borrowing and taboo, *Language*, **39**: 60–66.

OGAWA, N. and ASAI, E.
1935 *The Myths and Traditions of the Formosan Native Tribes.* Taihoku.

TUNG, T. H. (MS)
n.d. *A descriptive study of the Tsou language*, Pt. 3: Glossary.