BASES FOR NEW METHODS IN GLOTTOCHRONOLOGY

Jacques B.M. Guy

This paper was drafted on DEC-KL10 using VIDED, a display-oriented text editor written by Jacob Palme, of the Swedish National Defense Research Institute. All software referred to in the paper was written by the author in SIMULA, an ALGOL-based language developed by the Swedish National Defense Research Institute and the Norwegian Computing Center. Numbers in square brackets refer to figures and tables.

INTRODUCTION

This paper is at the same time a very condensed report of past work (Guy 1980) and an outline of work in progress which originated in a tentative clarification of the internal and external relationships of the languages of the New Hebrides contributed to the paper presented by Tryon at the Second International Conference on Austronesian Linguistics held in Canberra in 1978. To quote him:

> ...Guy (forthcoming) has proposed that a new technique involving linear regression might well be applied and prove more worthwhile than the lexicostatistical techniques that have become traditional. The technique has been applied to the percentages computed for the languages considered in this paper, and appears as Chart IV. (895)

The idea of seeking entirely new methods for classifying languages had stemmed from an earlier dissatisfaction with the results produced by traditional lexicostatistics. Having collected 22 of the 40 wordlists representing Santo in Tryon's internal classification of the New Hebrides languages (Tryon 1976), and become fluent in the two Sakao dialects and in Tolomako, I could only strongly disagree with techniques that set up an East Santo group containing Sakao as distinct from the North and Central New Hebrides as Tanna is. To me, Sakao was clearly a Santo language, although I would have been at great pains to substantiate this purely impressionistic judgement.

Possibly in this case the method was sound but, owing to the sheer amount of data produced (15,753 cognate percentages), it had not been possible to apply it properly.

A survey of the literature on the practice and theory of lexicostatistical methods, far from providing any solutions, suggested that the methods in current use rested on unsound foundations. In 1962 Bergsland and Vogt presented conclusive evidence that retention rates could vary greatly across languages. In 1965 Dyen, James and Cole had calculated the retention rates of the individual items of a standard sample wordlist, which were also found to vary.

Amran Halim, Lois Carrington and S.A. Wurm, eds *Papers from the Third International Conference on Austronesian Linguistics,* vol.1: *Currents in Oceanic,* 283-314. *Pacific Linguistics,* C-74, 1982.

THE ASSUMPTIONS OF GLOTTOCHRONOLOGY EXAMINED

Glottochronology is based on a hypothesis expressed by the formula

$$t = \frac{\log c}{2 \log r} \qquad (1)$$

where c is the proportion of cognates between two languages which have split t units of time ago and have since evolved at such a rate that any item of the sample word-list had a chance r (such that $0<r<1$) to survive, unreplaced, through a unit period of time.

To quote Swadesh:

> The mathematical formula is
>
> $$i = \frac{\log C}{2 \log r}$$
>
> in which i stands for indicated periods of depth, C is the percentage of basic vocabulary in common and r is the percentage of basic vocabulary retained after one standard period of time. The basic theory leading to this formula is presented in the next section.
> [...]
> 2.1. Assuming a constant average rate of drift in the basic vocabulary, the effect of elapsing time on an original vocabulary will be to reduce the vocabulary by a given percentage at the beginning of that period. If 85% is left after one period of time, then there will be 85% of 85%, or 72%, after a second equal period. After the third period there will remain 85% or 72% (that is 61%) of the original and so on for each additional period. (Swadesh 1950:158)

Swadesh's basic theory is expressed in somewhat obscure terms. What is a constant average rate? Swadesh obviously did not mean a constant rate, or he would not have used the word 'average'. An analogy may provide a hint of what he meant:

> 'we drove from **** to **** at a constant average speed of 90km/h'

At first this sentence seems to contain a contradiction in terms, but, on further reflection, one possible meaning emerges:

> 'we drove from **** to ****, calculating our average speed on several legs of the trip; we got an average of 90km/h every time.'

Which prompts the comment:

> 'either you drove at a constant speed, or you decided to call it the end of a particular leg of your trip when you had calculated that your average speed would be precisely 90km/h for that leg, or this is a very extra-ordinary coincidence indeed.'

It is quite unreasonable to hold that language families allow their members to split only at those times when their average rates of drift happen to be equal to a certain constant value. Since Swadesh did not hold that rates of drift were constant one can only conclude that his fundamental assumption is that basic vocabulary drifts (i.e. gets replaced) at a rate which fluctuates within narrow, constant limits.

Consider now a case where the retention rate is 90% for one half of the basic vocabulary and 10% for the other. After one period of time 90% of one half will be left, and 10% of the other half, i.e. 50% of the total basic vocabulary. You might expect 50% of 50%, or 25%, to be left after two periods of time. Not at all. 90%

of 90%, or 81%, of one half will be left, and 10% of 10%, or 1%, of the other half, i.e. 41% of the total basic vocabulary. And after three periods of time 36.5% will be left, not 12.5%. A second assumption, that the rate of drift is the same (or about the same) for all parts of the basic vocabulary, has to be made.

The findings of Bergsland and Vogt and of Dyen, James and Cole belie both assumptions.

Import of Bergsland and Vogt's findings

The variations which Bergsland and Vogt have observed in the retention rates of closely related languages or dialects (East Greenlandic and West Greenlandic for instance) are so great that, even assuming that the probability of survival of individual items is equal throughout a given wordlist, time depths calculated according to formula (1) are subject to such large errors that they become utterly useless. Thus for instance, if told that languages X and Y split 1000 years ago, relying on a retention rate of 86% per millennium (the one generally used for the Swadesh 100-item wordlist), one would confidently predict that their most likely percentage of shared cognates should be around 73.96%. If however X and Y had had the bad taste of evolving at the same rates as East and West Greenlandic (53.7% and 97.2% retention per millennium respectively), they would most probably share 52.2% cognates. Neither would the practice of calculating a range of error for a 95% level of confidence help in the least: on a 100-item wordlist 73.96% and 52.2% are just about four standard deviations apart, which would require a confidence level of over 99.99%.

Bergsland and Vogt's study, which showed beyond reasonable doubt that a percentage of shared cognates is not an accurate enough measure of the genetic distance between two languages, should have brought about drastic, if painful, revisions in the lexicostatistical method. It was ignored.

Import of Dyen, James, and Cole's findings

Any linguist who has collected or perused sample wordlists from Austronesian languages is bound to have noticed that certain words (e.g. 'two', 'five', 'eye', 'die', etc.) are extremely stable throughout Austronesia.

Dyen, James, and Cole have calculated the retention rates (which should more accurately be called chances of survival) for 196 items of a sample wordlist from 46 pairs of Austronesian languages. In the table they give (Dyen 1967:153-154), chances of survival are not listed as such, but plodding through the text eventually shows that the chance of survival of an item through 1000 years can be calculated from the formula:

$$r = e^{-\frac{1}{2.1382\tau}} \qquad (2)$$

where e is the base of the natural logarithms (2.78128...), and τ a time value the inverse of which is given for each item of the wordlist.

Survival chances thus extend from 0.9918 per thousand years for 'five' down to 0.5857 for 'to play'. Their values are normally distributed, without any sign of skewness, kurtosis, or any other irregularities, with a mean value of 0.7916 (i.e. 79.16% survival chance through 1000 years) and a standard deviation of 0.0942.

The standard deviation is an arithmetic quantity which expresses how far from their mean a set of values spreads. Here the spread is about three times as large as could be attributed to chance alone.

How does this affect the fundamental formula (1) of glottochronology?

Consider a wordlist consisting of n items, with survival chances r_1, r_2, r_3, ...r_n for language L and r'_1, r'_2, r'_3, ...r'_n for language L'. If L and L' have split t time units ago the most probable proportion of shared cognates (call it c) between them is given by the formula

$$c = \frac{1}{n} (r_1^t r'^t_1 + r_2^t r'^t_2 + r_3^t r'^t_3 + \ldots r_n^t r'^t_n) \tag{3}$$

or, in more compact notation:

$$c = \frac{1}{n} \Sigma (rr')^t \tag{3a}$$

It is only in the very special case where all items have the same chance of survival that the above formula reduces to

$$c = \frac{1}{n} (r_1^{2t} + r_2^{2t} + r_3^{2t} + \ldots r_n^{2t}) = \frac{1}{n} nr^{2t} = r^{2t} \tag{4}$$

from which follows

$$t = \frac{\log c}{2 \log r}$$

Disregarding the findings of Bergsland and Vogt and working on the assumption that the chance of survival of any one item of the wordlist is the same for all languages, the general formula (3) becomes

$$c = \frac{1}{n} (r_1^{2t} + r_2^{2t} + \ldots r_n^{2t}) = \frac{1}{n} \Sigma r^{2t} \tag{5}$$

But, under the false hypothesis that the survival chances of individual items are equal, we would calculate c as

$$c = p^2 \tag{6}$$

where p is the proportion of vocabulary each of the two lists is most likely to have retained since the split of the languages they represent, and is given by

$$p = \frac{1}{n} (r_1^t + r_2^t + r_3^t + \ldots r_n^t) \tag{7}$$

From (6) and (7) we have

$$c = \frac{1}{n^2} (r^t + r^t + \ldots r^t)^2 = \frac{1}{n^2} (\Sigma r^t)^2 \tag{8}$$

The error on the calculation of c is the difference between (5) and (8)

$$\frac{1}{n} \Sigma r^{2t} - \frac{1}{n^2} (\Sigma r^t)^2 \tag{9}$$

which happens to be the square of the standard deviation of r^t.

From Dyen 1967 we know that, on his 196-item list, with t equal to 1000 years, and the values of r (i.e. the individual chances of survival) observed, the standard deviation of r^t is 0.0942. Told that languages X and Y split 1000 years ago, one would therefore estimate their most likely proportion of shared cognates at present as

$$0.7916^{2t} = 0.6266, \text{ i.e. } 62.66\%$$

whereas in actual fact their most likely proportion of shared cognates would be

$$0.7916^{2t} + 0.0942^2 = 0.6355, \text{ i.e. } 63.55\%$$

a negligible error of less than one percentage point.

For greater time depths, however, the error increases sizably, as the standard deviation of r^t is given by

$$\sigma \simeq \frac{(0.7916 + 0.0942)^t - (0.7916 - 0.0942)^t}{2} \qquad (10)$$

Thus for instance:

Time depths	Percentages of shared cognates predicted by formula (8)	by formula (5)
2000 years	39.3	41.5
3000 years	24.6	27.8
4000 years	15.2	19.0
5000 years	9.7	13.3

The error introduced by disregarding the fact that items of a sample wordlist do not usually stand equal chances of survival remains minimal compared to the enormous misevaluations of time depths caused by ignoring Bergsland and Vogt's findings.

GLOTTOCHRONOLOGY (ALMOST) WITHOUT RETENTION RATES

Abstracting the effects of Bergsland and Vogt's findings

Imagine a language L splitting into communalects X and Y ([1]). One thousand years later, X has retained 90% of its vocabulary whereas Y, for some reason or other, has retained only 30%. X and Y are the only two surviving descendants of L, but have many 'cousins' (parent languages derived from an ancestor of L). X will score on the average three times as many cognates with any cousin as Y with this same cousin (you are urged to verify this statement by doing the calculation yourself). Let us now draw a graph ([2]) where each of X and Y's cousin languages is represented by a point the co-ordinates of which are its percentages of shared cognates with A1 and A2, so that if a language Z has 42% cognates with X and 13% with Y it is represented by a point 42mm from the X axis and 13mm from the Y axis. The percentage score x of any cousin language with X being on the average 3 times its score y with language Y, these points will tend to be found along the straight line corresponding to the linear equation

$$y = \frac{1}{3}x$$

Now let Y split into a number of communalects Y1, Y2, etc., which start replacing vocabulary at different rates. Assuming that all survive, consider the situation after 1000 years ([3]). The scores of any of Y's descendants with its cousins (B, C, D, ... Z) will still tend to be in a constant ratio with the scores of X's descendant with those same cousins, *whatever their individual retention rates* for the past 1000 years. If their scores with X's descendant and any given descendant of Y are again plotted as in [2], cousin languages will now again be represented by points closely clustered along a straight line, yielding a high, positive linear correlation co-efficient; if now plotted in the same manner onto the same graph, the sister languages of Y1 (Y2, Y3, etc.) will tend to be scattered off that main line or to cluster along secondary lines. Whichever the case, their addition to the graph obscures the single-line pattern and the linear-correlation co-efficient drops accordingly. The lower correlation between X and Y1 is brought about by the fact that they no longer have a common immediate ancestor. Thus high linear-correlation co-efficients point to the most recent splits of all, whatever the individual past retention rates.

> An algebraic quantity called linear correlation co-efficient expresses how close to a line a set of points clusters. Its sign indicates in which direction the line slopes and its absolute value can vary from 1 (all points exactly on the line) down to 0 (no linear pattern at all). The basis for its computation is the square of the distance of each point to the line along which they all appear to cluster. The computational short cut for finding the line that passes closest to all points, such that the sum of the squares of their distances to it is lower than it would be for any other straight line, is called linear regression.

Graph [4] shows the proportions of cognates for 176 lists from the New Hebrides (taken from Tryon 1976) plotted against Sakao (list 64) and Shark Bay (list 67). Sakao and Shark Bay belong to an East Santo group (Tryon's classification, 1976:86-87). Note how the overwhelming majority of points cluster about line OA, and how the five points representing the other five lists taken from the East Santo languages cluster off the main line, about a nearly horizontal line, in the Shark Bay half of the diagram. The Sakao half of the diagram, on the other hand, is empty: list 64 (Sakao, Port Olry dialect) is the only representative of the Sakao subgroup.

Graph [5] shows again the proportions of cognates for 176 lists plotted this time against Sowa (list 33, south Pentecost) and Hukua (list 34, north-west Santo). Sowa and Hukua are not closely related (Tryon assigns them to two different subgroups — West Santo and East New Hebrides — within the same North and Central New Hebrides group). Note how the majority of the dots now swarm within a roughly triangular area, with eight dots clearly outside that area: Seke, Apma, and Sa (three languages and dialects from Pentecost) on the Sowa side, and Piamatsina, Vunapu, Tasmate, Nokuku, and Valpei (five languages and dialects from north-west Santo) on the Hukua side.

Graph [6] shows all non-Santo lists plotted against Akei (south-west Santo) and Sakao. Here again a clear linear pattern shows up; the spread of the dots representing the non-Santo lists off line OA is about the same as that of the non-East Santo lists in graph [4]. In other words, Sakao and Akei are as closely related within Santo as Sakao and Shark Bay are within East Santo. Line OA in diagram [6] is quite steep: its slope is about 1.4, which means that the percentage of cognates between Akei and any related language X outside Santo is approximately equal to 1.4 times that of Sakao with X, or that the average resistance to change of Sakao since their separation has been only about 70% than that of Akei. According to Tryon Akei and Sakao have 33.55% basic vocabulary in common. Call S the proportion of vocabulary retained by Sakao from its immediate common ancestor with Akei, and A that retained by Akei. Their most likely proportion of cognates is then

$$c = S \times A = 0.7\ A^2 \quad \text{(since } S = 0.7\ A\text{)}$$

of which the value of 33.5% observed is the best (and only!) estimate we can hope to have. The most likely proportion of vocabulary retained by Akei since the split is therefore:

$$A = \sqrt{\frac{c}{0.7}} = 0.6918 \quad \text{i.e. about 69\%}$$

and by Sakao:

$$0.7 \times 0.6918 = 0.4843 \quad \text{i.e. about 48\%}$$

If assumed to have evolved at the rate generally accepted in glottochronology (81% vocabulary survival per millennium), Akei would then have separated from Sakao around 1750 B.P., whereas the standard formula of glottochronology would have given a date of 2595 B.P. It is not surprising that Tryon, relying on current lexicostatistical methods, should have been led to classify the East Santo languages in a group as distinct from the rest of Santo as are, for instance, the languages of Tanna.

Abstracting the effects of Dyen, James and Cole's findings

The error introduced by disregarding Dyen, James and Cole's findings is negligible for fairly high percentages of shared cognates, but increases for lower and lower percentages. As they show in their article, it is possible to obtain an estimate of the individual retention rates of the different items of a sample wordlist. From these, one can then calculate what the proportion of cognates between two wordlists would have been if all items had had the same retention rate. The procedure is within the capacity of a small modern computer even though it is quite complex, involving lengthy computations and iterations, and would have to be modified to take into account unequal overall retention rates across languages.

Since the error is so small in absolute terms, it may be possible to disregard it in practice. Better still, when computing proportions of shared cognates, one sould discard items with extremely high or extremely low retention rates: lexical innovations are the clues to past language splits; those items with extremely high retention rates, being so seldom replaced, provide little evidence (if any at all: it is a rare Austronesian language indeed which does not have some reflex of dua for *'two'* or mata for *'eye'*), whereas the evidence from items often replaced gets constantly overwritten by later innovations. Lexical items which show extreme retention rates are therefore devoid of information and are advantageously ignored.

CONCLUSION

Reliability of the method

Since the method rests on much less restrictive hypotheses than the traditional lexicostatistics, its results can only be as good or better than those of lexicostatistics as currently practised. This is not to say that they are absolutely reliable. Traditional lexicostatistical methods are rather unreliable (Guy 1980) even when applied to data produced under ideal conditions: no borrowings, no semantic shifts, and a uniform, constant retention rate, equal for all items and all lists. With hindsight, such failures are not surprising: in classifying languages, glottochronologists rely on very small samples indeed; after all, a 200-item wordlist does not represent even 1% of the active lexicon of a language. Lengthening the standard wordlist to a more substantial 1000 items would greatly increase the reliability of the classifications, although not fivefold. But the purpose of lexicostatistics applied to language taxonomy must be kept in mind: to provide a simple, quick method for classifying a large number of ill-known communalects. Advocating the use of 1000- or 2000-item wordlists runs directly against this aim.

Guidelines for practical applications

From the examples of scatter diagrams given thereabove (Sakao-Shark Bay, Sakao-Akei, Sowa-Hukua), it should be clear that the more lists from which percentages of shared cognates have been computed, the more reliable the method is.

It is difficult to give an absolute, practical minimum number of lists under which the method should not be applied. I certainly would not contemplate using it on just a dozen lists. Furthermore, the more distantly related the languages to be classified, the more should be represented in the sample. For instance, whereas I would feel reasonably confident about a classification of the New Hebrides languages from thirty well-chosen lists, I would see a classification of Austronesian or Indo-European based on thirty lists as seriously inadequate.

The calculation of correlation co-efficients is a tedious task, even with the

help of one of the many pocket calculators now available with inbuilt linear-regression facilities, all the more so that the method can be fruitfully applied only to large tables of percentages (the larger the table, the more diversified the percentages, the better). A computer should be used whenever possible, not so much to save time as to eliminate the element of human error due to fatigue and sheer boredom.

It is out of the question that one should draw (or even have drawn by computer) diagrams such as graphs [4], [5], and [6] for each language pair. But one should certainly not fail to do so for a few selected pairs, and preferably by hand: this very exercise somehow seems to bring one valuable insights.

Having obtained a table of linear-correlation co-efficients and a few graphs from a table of percentages of shared cognates, one is still left with the task of interpreting this new set of data. Low-level groupings will be immediately evident from the table of correlations, but high-level ones are normally far from obvious. Several methods are developed and tested in Guy 1980, one of which recommends itself by its greater reliability. Although the calculations it requires can be carried by hand, the use of a computer is recommended, again to avoid human errors. An alternative method, highly commendable for its speed and ease of execution, is the single-linkage algorithm (Hartigan 1974:193-194). Unfortunately, it is not specifically designed to produce phylogenetic reconstructions.

A wider viewpoint and some critiques

The classification of languages by lexicostatistical methods is but a special case of a much more general problem which can be expressed in these terms:

A message is input at the root of a tree-shaped communication network, from which it travels forward to the leaves. The further it travels, the greater its chances of being distorted by noise in the network. From the garbled versions of the original message collected at the leaves, reconstitute the network.

In the model used in this paper, the message was a string of symbols; the set of all possible symbols — the alphabet of the message — was the set of all possible word shapes in the proto-language; there was only one kind of distortion: a symbol affected by noise being replaced by a symbol randomly chosen from the alphabet of the message; the probability of a symbol being affected by noise at any given point of the network was proportional to the product of the amount of noise at that point by its stability, the latter being a function of the meaning of the symbol (i.e. its position in the message); finally, the alphabet of the message was taken to be infinite.

The model presents the following discrepancies with reality:

The alphabet of the message is not infinite. No language may draw from an infinite vocabulary, unless its number of phonemes is infinite, or its possesses words of infinite length, both impossible conditions. But it is very large, so that the probability of a symbol being replaced by itself or by a symbol which happens to occur in the same position in some other version of the message somewhere else in the network, is — hopefully — low enough to constitute only a minor disturbance.

The stability of a symbol is not necessarily solely a function of its position in the message (i.e. of its meaning). I, for one, would rather see it also as a function of phonological decay resulting in an annoyingly large number of homophones.

Distortions of the message can occur in other ways than in the model: semantics swaps (symbols of the message exchanging positions), borrowing or convergence

(interferences, cross-talk, poor channel separation). When these play a major role, the model does not apply, and it is especially to be hoped that the method presented will never be used in a tentative classification of Interlingua, Esperanto, Volapul, Pidgin English, and the Chinook Jargon. When they play a minor role, the reliability of the method is only reduced (certain algorithms, based on the direct examination of the wordlists reduced to reconstructed proto-forms, stop in the presence of false cognates such as may arise from accidental cognate-like forms, borrowings, semantic swaps, etc., and may some day provide a solution to this problem).

APPENDIX: EXAMPLES OF APPLICATION

Ideally, the method should be tested on data from language families the evolution of which has been ascertained beyond reasonable doubt on the basis of extralinguistic evidence such as written history and archaeological findings.

Unfortunately, there are no published tables of shared-cognate percentages for such language families (such as Indo-European) that are based on enough wordlists for the method to be applicable.

New Hebrides

The only set of data involving an undoubtedly large enough number of wordlists was published in Tryon 1976 in his classification of the New Hebrides languages. Having written the computer program which was to produce those sixty pages of cognate percentages, collected a number of wordlists from Santo, and attempted reconstructions of Proto-New Hebridean from languages of Santo, Aoba, and the Banks, I could hardly resist the temptation of testing the method on this set of data, even though next to nothing is known about the prehistory of the New Hebrides, and even though it involved entering 15,743 cognate percentages by hand, since, despairing of ever being able to have them analysed automatically, I had rather stupidly overwritten the tape containing the original data. Although no correction could be applied to counterbalance the effects of unequal item retention rates (as the necessary coded wordlists were erased when the original tape was overwritten), judging from the nature of the data and of the algorithms used, I do not believe that the classification would have been any different, had the proper corrections been applied.

The classification obtained is vastly different from that achieved by current lexicostatistic methods (Tryon 1976, 1978). Table [7] gives the output from the *n*-way splitting algorithm described in Guy 1980, applied to the linear correlation co-efficients computed from Tryon's percentages. Each wordlist is identified by the name and number (the latter in parentheses) used by Tryon, and is followed by a string of digits showing to which successive groups and subgroups it was assigned by the algorithm. The lists are arranged in such an order that lists belonging to the same subgroup are found next to each other. The corresponding genealogical tree was added by hand.

According to this classification the languages of the New Hebrides fall into four great groups.

Group 1, call it North New Hebridean, encompasses the Banks and Torres islands, Santo and its offshore islands, Aoba, Maewo, and Pentecost.

Group 2, call it Central New Hebridean, contains Malekula and its offshore islands, Ambrym, and Epi.

Group 3, call it Hebrido-Polynesian, comprises the Shepherd Islands, Efate, and the Polynesian Outliers (Makatea in the Shepherd Islands, Fila-Mele on Efate, and Futuna and Aniwa islands much further south).

Group 4, call it South New Hebridean, consists of the three southernmost islands, Erromanga, Tanna, and Aneityum.

These groups subdivide as follows:

North New Hebridean
- A. North-east New Hebridean
  - 1. Northern subgroup
    - a. Torres (Hiw, Toga)
    - b. Motan (Ureparapara, Mota, Motlav, Vanua-Lava)
    - c. Gauan (Gaua, Merlav, Merig)
  - 2. Southern subgroup
    - a. Maewo
    - b. Aoban (Aoba and northernmost tip of Pentecost)
    - c. Pentecostan (rest of Pentecost)
- B. Santo
  - 1. North-west Santo (Valpei, Nokuku, Hukua, Vunapu, Piamatsina, Tasmate)
  - 2. North-east Santo
    - a. Sakao
    - b. Shark Bay
    - c. Vanafo-Luganville (the Butmas and Tur dialects have long been spoken in Vanafo, and Polonombauk village is just north of Luganville)
  - 3. South Santo
    - a. South-west Santo
      - West Coast (Wusi with an inland enclave: Nonona)
      - North Coast (Tolomako, spoken from south of Piamatsina to the mouth of the Jordan river)
      - South Coast and Inland (includes a very aberrant language: Roria)
    - b. South-east Santo
      - Mainland
      - Offshore Islands (including a mainland enclave: Tambotalo)

Central New Hebridean
- A. Malekula
  - 1. North Malekula
    - a. Big Nambas (Maragus, Unmet, and Leviamp)
    - b. East Coast Islands (Atchin, Wala, Rano, Uripiv, Uri, with two mainland villages: Pinalum, Tautu)
    - c. North Coast (north of Big Nambas, and Vao island)
  - 2. South Malekula
    - a. South-east Coast (the southern half of the east coast of Malekula, and the eastern half of its south coast, with the offshore islands)
    - b. Mainland (the rest of south Malekula, Windua and Labo excepted)
    - c. Windua and Labo villages
- B. Ambrymese
  - 1. Mainland (all Ambrym except Maat and Toak)
  - 2. Offshore (Paama with, on south-east Ambrym, Maat and Toak, and on east Epi, Laul)
- C. Epi
  - 1. South-west Epi (two communities: Mae-Morae and Vowa)

2. Main
   a. East Epi (one language, Lewo, with dialect chaining)
   b. West Epi (two languages, Bierebo and Baki)

Hebrido-Polynesian

A. New Hebridean Component
   1. Namakura (one language spoken in the Shepherd islands)
   2. Efatese (on Efate and parts of the Shepherd islands)

B. Polynesian Component (Makatea in the Shepherd Islands, Fila-Mele on Efate, and Aniwa-Futuna much further south, east of Tanna)

South New Hebridean

A. East Erromangan (Ura)

B. West Erromangan (Sie)

C. Aneityum

D. Tanna
   1. North Tanna
      a. North-east Tanna
      b. North-west Tanna
   2. South Tanna
      a. East Coast (Port Resolution, Yatukwey, and Iarkei)
      b. Mainland (rest of the southern half of Tanna)

On the hypothesis that areas of greater linguistic diversity tend to betray centres of diffusion, a tentative map of prehistoric migrations can be drawn [8]. The enterprise is not without its pitfalls, for it is sometimes difficult to decide where the area of greater linguistic diversity lies. Within the southern subgroup of North-east New Hebridean for instance, there seems to be no clear possible choice between Aoba or Maewo, or even Pentecost. However, for North New Hebridean the area of greatest linguistic diversity is clearly Santo (and for Santo, south Santo), for Central New Hebridean it is Malekula (and probably, for Malekula, south Malekula), and for South New Hebridean, it is Tanna. The migrations thus reconstructed within North and Central New Hebridean suggests that the present-day populations of the New Hebrides came from the west. On the information provided by Tryon's percentages of shared cognates alone, there is no way of telling that the so-called Polynesian Outliers do not belong in the New Hebrides. The fact that the method classified them as a subgroup of Hebrido-Polynesian, rather than as a fifth group on the same footing as, say, North New Hebridean, suggests three possible interpretations:

1. Extensive, undetected borrowing has taken place between the New Hebridean and the Polynesian components. Which way the borrowing mainly went could only be discovered, if at all, by processing a table of proportions of shared cognates between a number of New Hebridean and Polynesian lists, including those Hebrido-Polynesian lists.

2. The New Hebridean component is an extremely aberrant branch of Polynesian.

3. The Polynesian component is an extremely aberrant branch of New Hebridean.

I do not consider the third hypothesis to have the slightest chance of being true, and the second hypothesis seems somewhat far-fetched. It is therefore most likely that extensive undetected borrowing did take place within Hebrido-Polynesian. Before a valid grouping of the New Hebridean component can be obtained, Polynesian loanwords should therefore be identified in the lists, the affected items discarded, proportions of shared cognates computed afresh, and the new table reprocessed. The New Hebridean component might then turn out to have joined one of the other three great groups or to have remained on its own.

Even in the absence of reliable information on the prehistory of the New Hebrides, it is possible to make a few non-trivial remarks.

The oft encountered subgrouping into offshore islands (usually with mainland enclaves) against mainland corresponds to a real settlement pattern: the inhabitants of offshore islands would, and in many cases (e.g. Vao and Wala) still do, keep gardens on the mainland.

The northward migration pattern postulated for Santo is corroborated by Sakao toponymy, tradition, and semantics.

The Sakao word for *'to go south'*, is a reflex of Proto-North New Hebridean ŋwule (Guy 1978), reflexes of which are found in the Banks (Lehali, Mota), on Aoba (Lolopuepue, Lolowai, Nangire), and on Santo (Tolomako, Tangoa) everywhere with the meaning *'to return home'*.

A Sakao tradition claims that Araki island (off the south Santo coast, opposite Wailapa) 'fled south', followed by all the other offshore islands (Tangoa, Malo, etc. up to Thion, right opposite Port Olry); this suggests a seaborne migration along the Santo coast and a settlement beyond Thion, very probably on Sakao island itself.

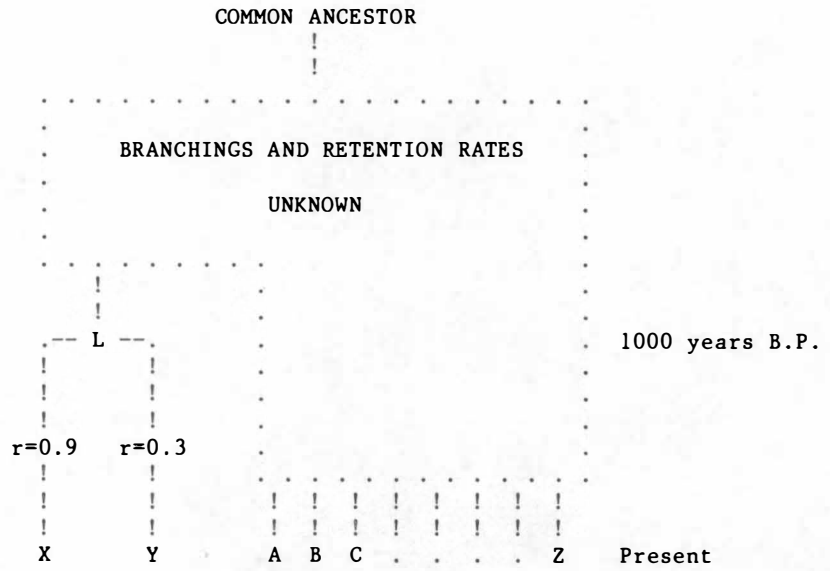
Finally Araki is the only island off the south coast of Santo to have, in Sakao, a proper Sakao name (Reg), regularly derivable from 'Araki', whereas all other islands have names (Aure, Malo, Tutupa, Ais, Mavea) alien to Sakao phonology and obvious, very recent borrowings; the name Araki must therefore have been part of the Sakao toponymy since a very early time, i.e. before the first consonant shift of Sakao (Guy 1978:790), which strongly suggests that the ancestors of present-day Sakao speakers left their original settlements on Araki and on the nearby mainland at a time when their communalect had barely diverged from Proto-North New Hebridean.

Polynesia

A number of tables of proportions of shared cognates have been published for the Polynesian languages. These tables exhibit a striking amount of disagreement between authors (thus Elbert gives 73% cognates between Maori and Tahitian, Dyen and Biggs 40.5% and 41% respectively), as different criteria of cognation were used (Elbert used the multiple-cognate, Dyen and Biggs the single-cognate method). The difficulty is compounded by the fact that no two tables involve the same languages, and that one (Kirk and Epling 1972:22) is composite, in part averaged from Dyen's, Elbert's, and Emory's counts, and in part recomputed to obtain best-distance estimates. Finally, the tables involve relatively few languages and dialects and thus the classifications which can be obtained by the linear-correlation method are all very much open to question. Nevertheless, the corresponding language groupings obtained using the *n*-way splitting algorithm are given in diagrams [9] to [13].

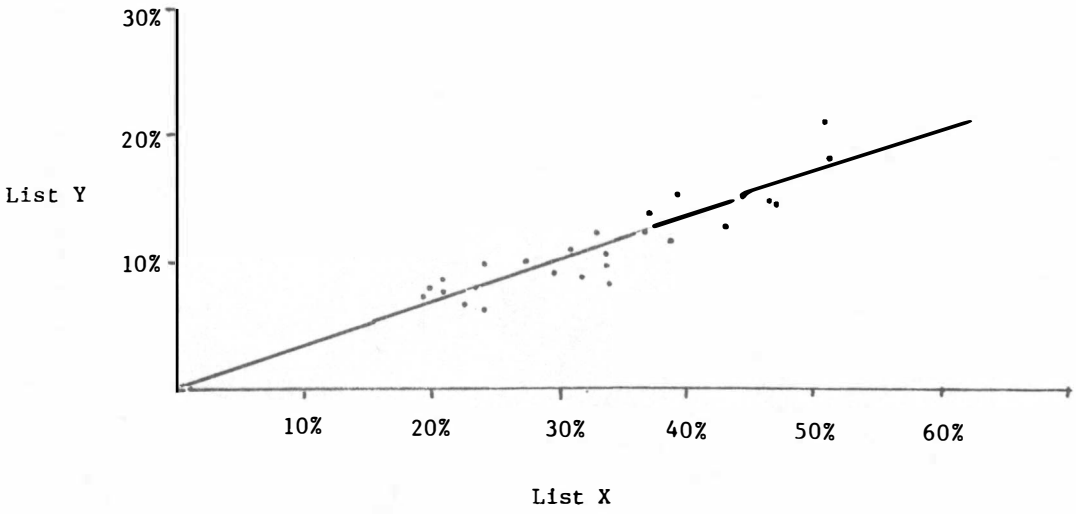
Twenty-five Austronesian languages

The figures are taken from Tryon 1978. Diagram [14] gives groupings obtained. Even though 25 lists are far too few for the method to be reliably applied, it can be pointed out how much better the groupings are than those obtained by current lexicostatistical methods (compare with Tryon 1978:892,894): not only do all the languages of the New Hebrides now form a distinct group on their own, with the three southernmost languages in a separate subgroup, but Xaraci (New Caledonia) and Nengone (Loyalty Islands) are also grouped together.



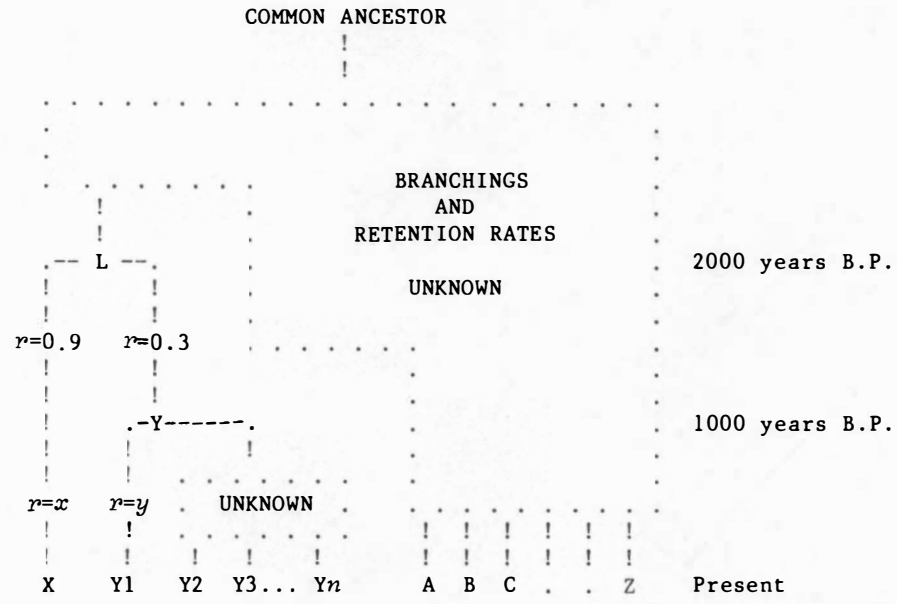
[1]

Whatever the branchings and the retention rates in the black box the proportion of cognates shared by X with any cousin language (A, B, C,... Z) will be around three times that of Y with that language.



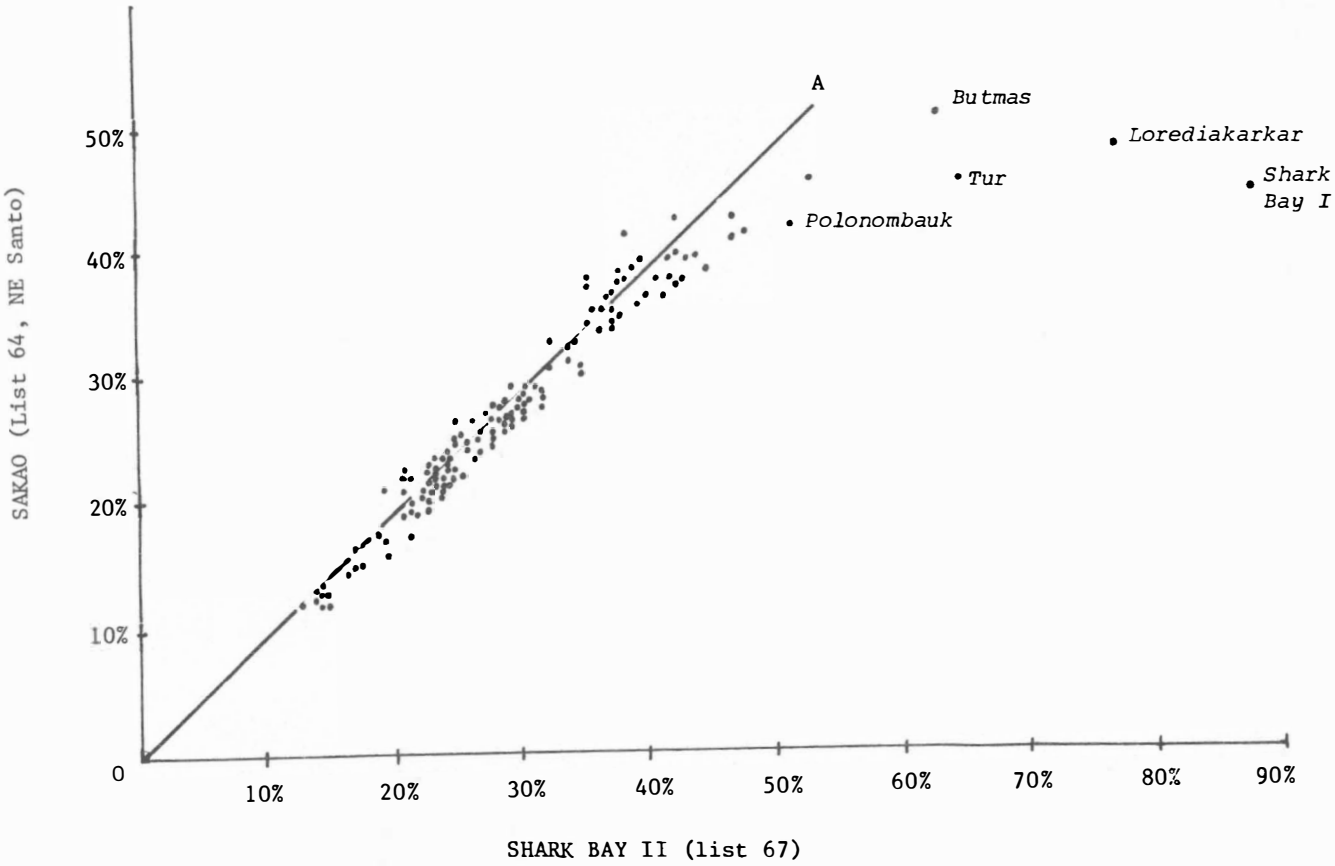
[2]

(See text p.287)



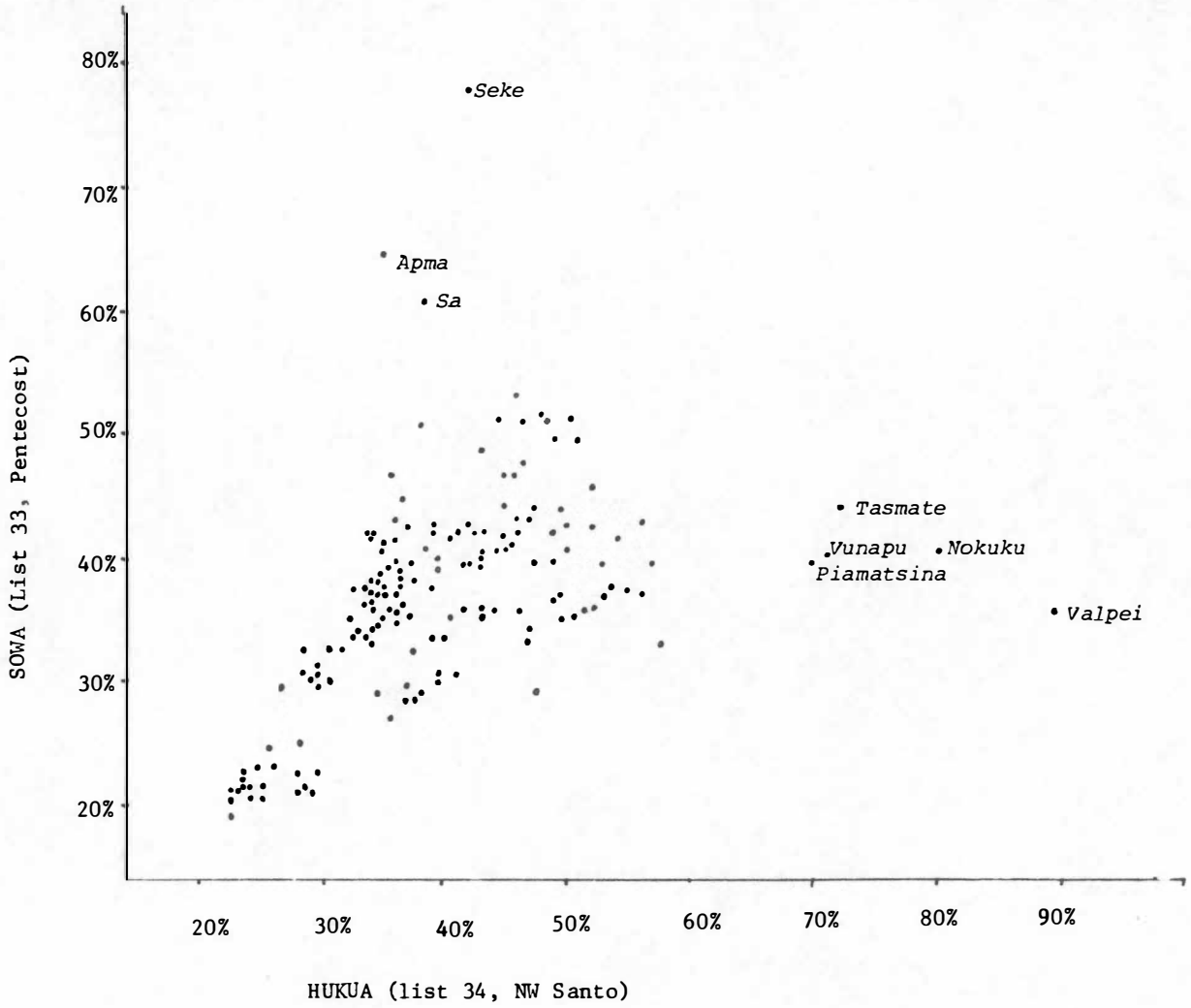
[3]

Whatever the branchings and the retention rates in the black box the proportion of cognates shared by X with any cousin language (A, B, C,... Z) will be around $3x/y$ times that of Y1 with that language. But the proportion of cognates shared by X with any of the other descendants of Y (Y2, Y3,... Yn) will be a variable ratio of the proportion of cognates shared by Y1 with that language, which ratio will depend on the branchings and retention rates between Y and its other descendants Y2, Y3,... Yn.



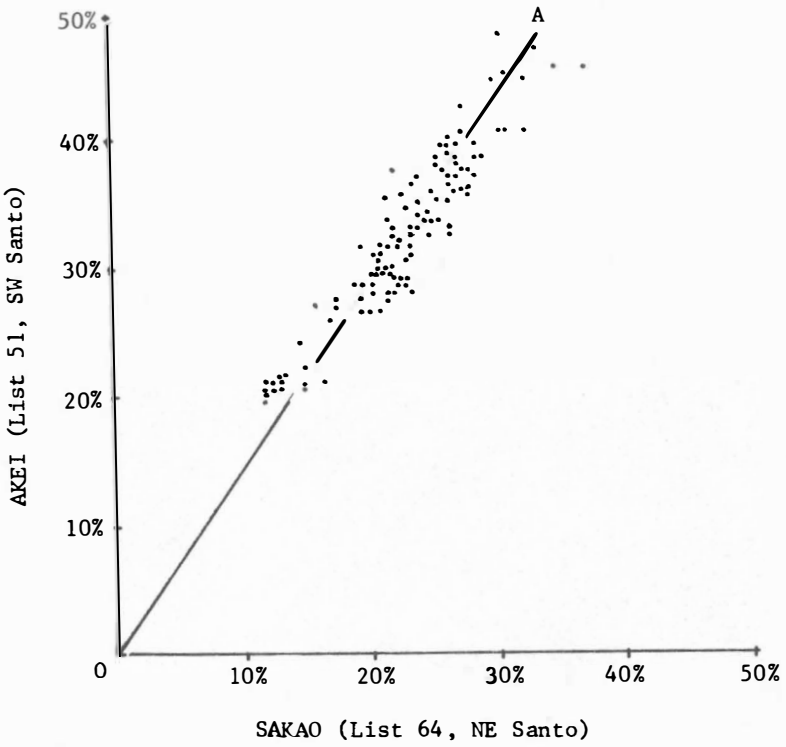
[4]

New Hebrides languages and dialects plotted against Sakao and Shark Bay



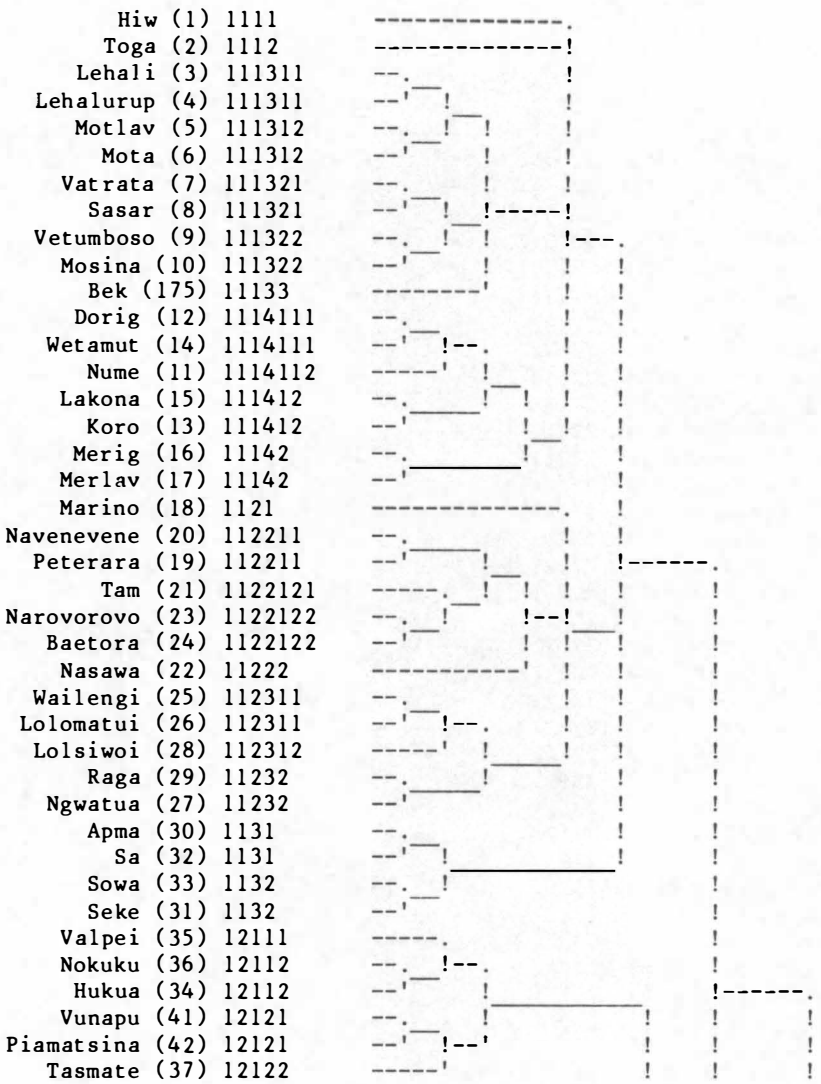
[5]

New Hebrides languages and dialects plotted against Sowa and Hukua



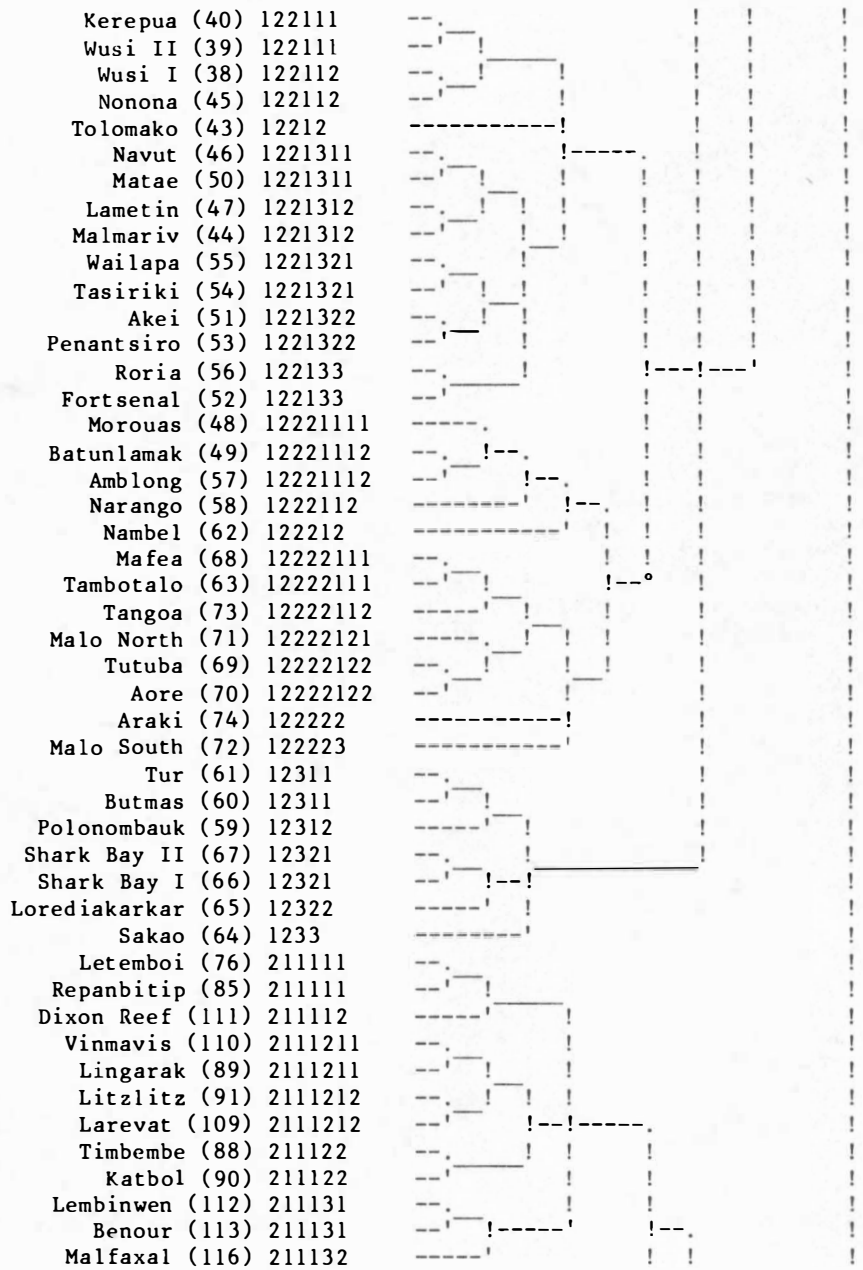
[6]

New Hebrides languages and dialects outside Santo plotted against Sakao and Akei



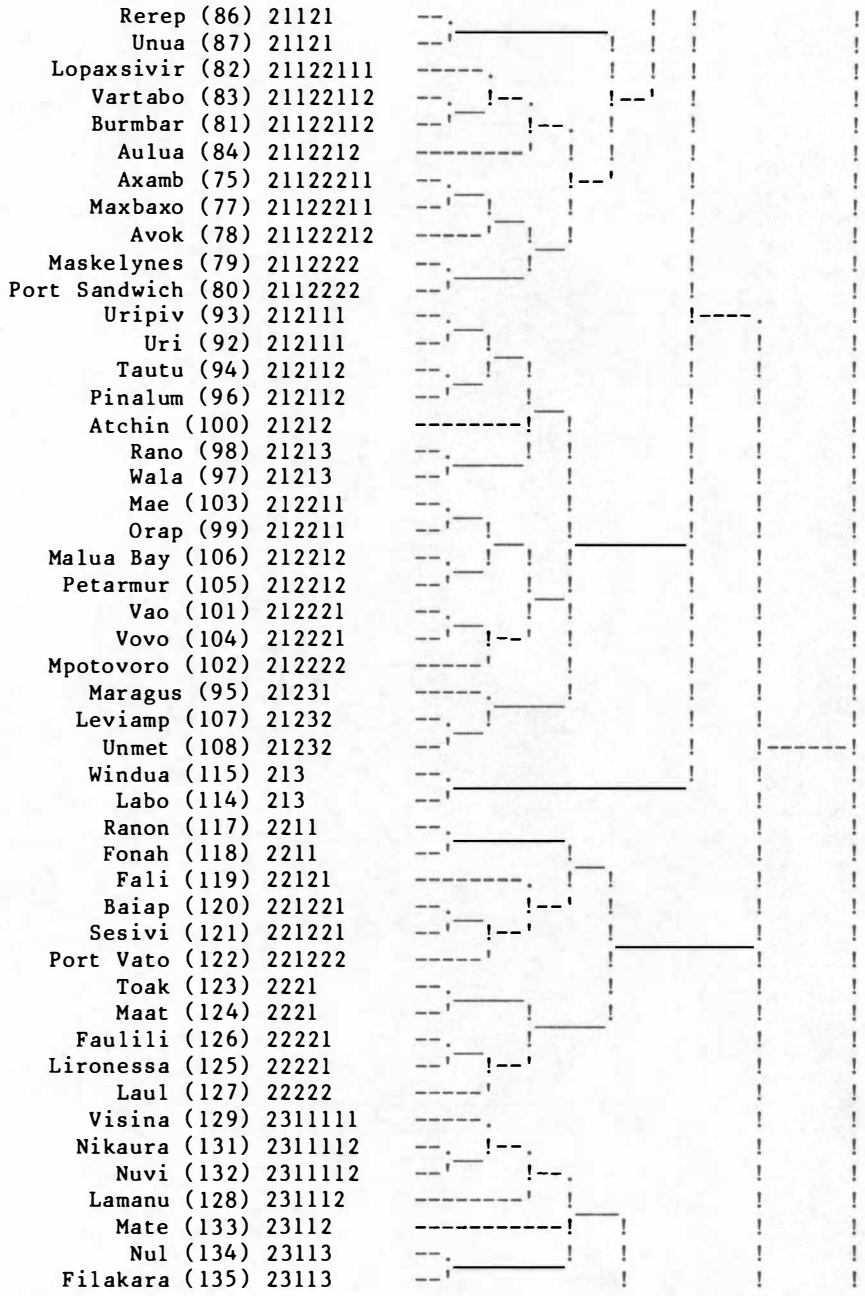
[7.1]

Reconstructed genealogy of the New Hebrides languages and dialects (cont...)



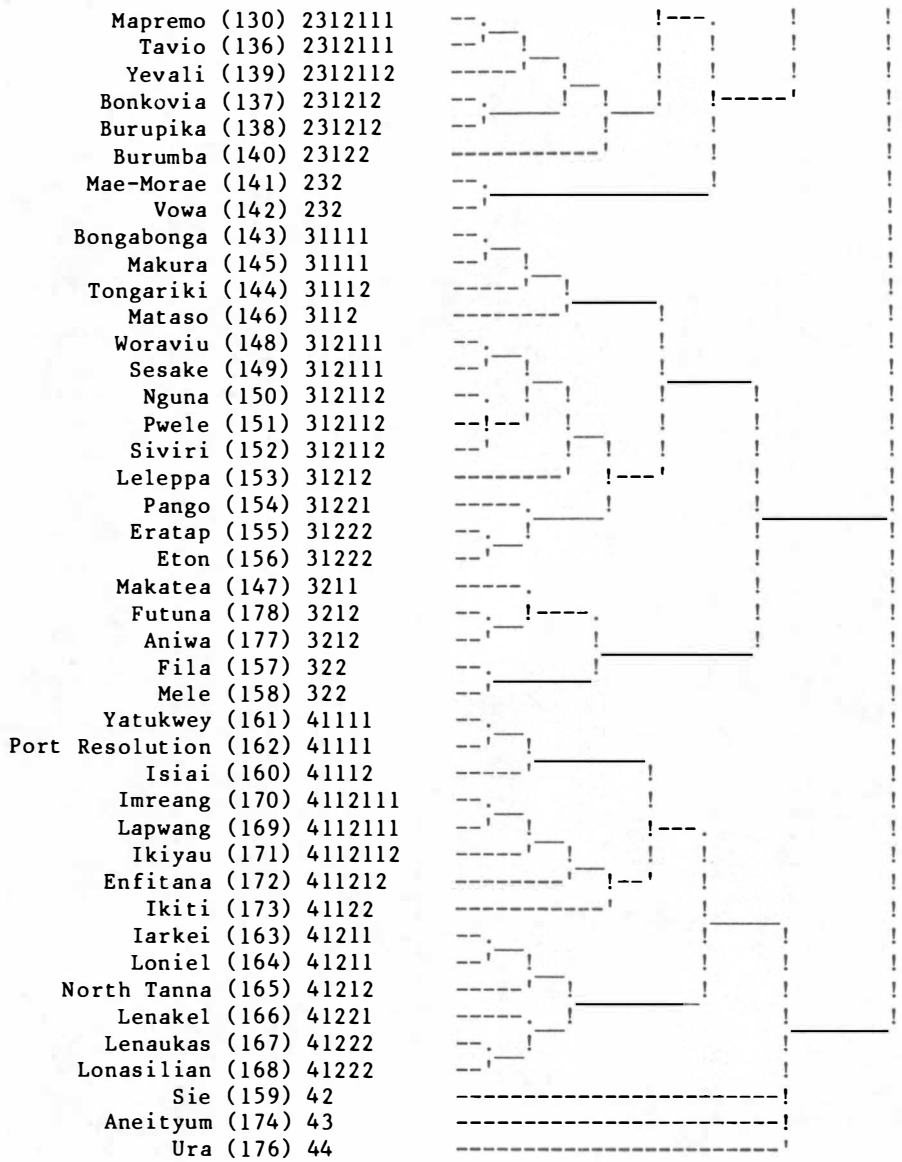
[7.2]

Reconstructed genealogy of the New Hebrides languages and dialects (cont...)



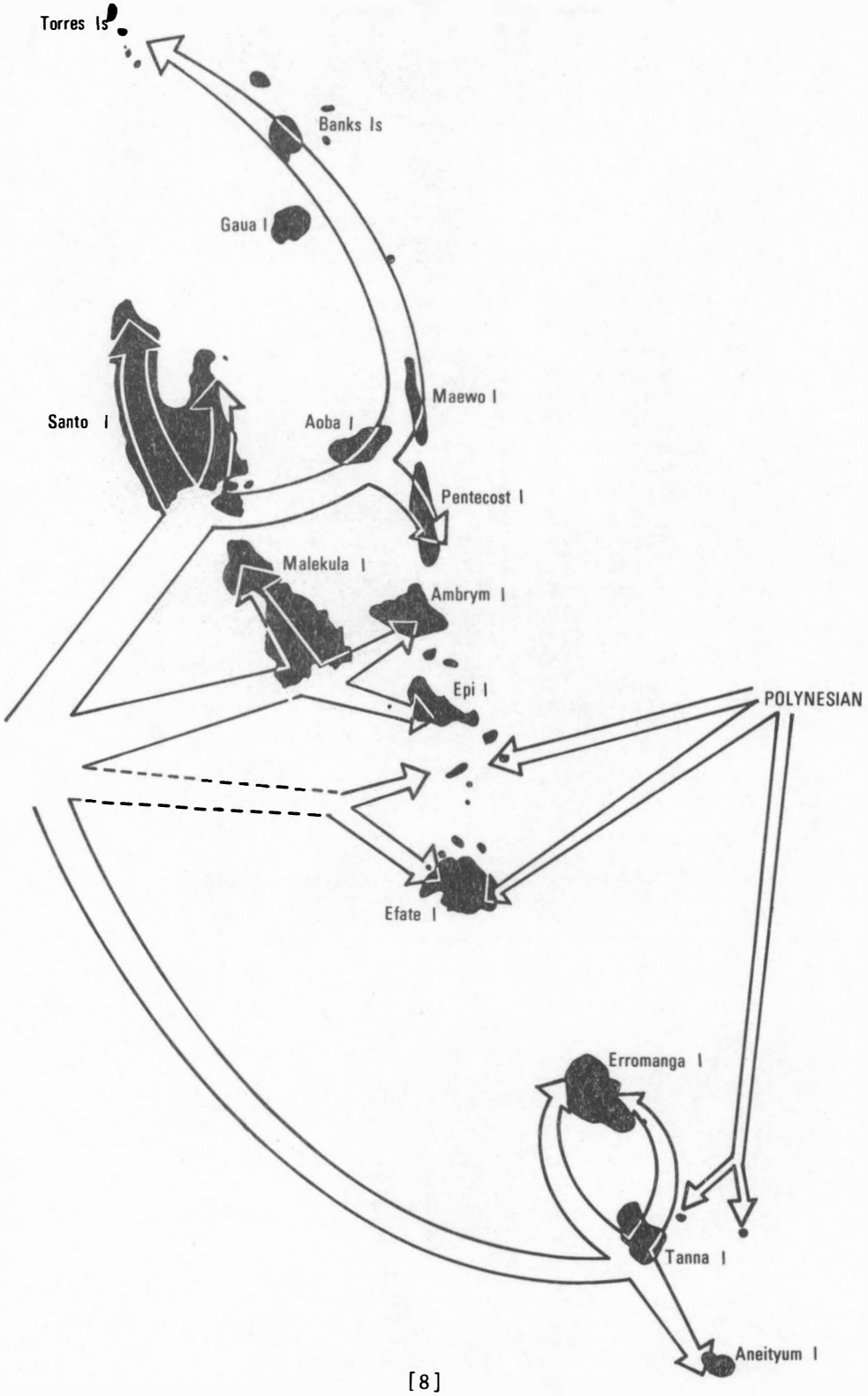
[7.3]

Reconstructed genealogy of the New Hebrides languages and dialects (cont...)



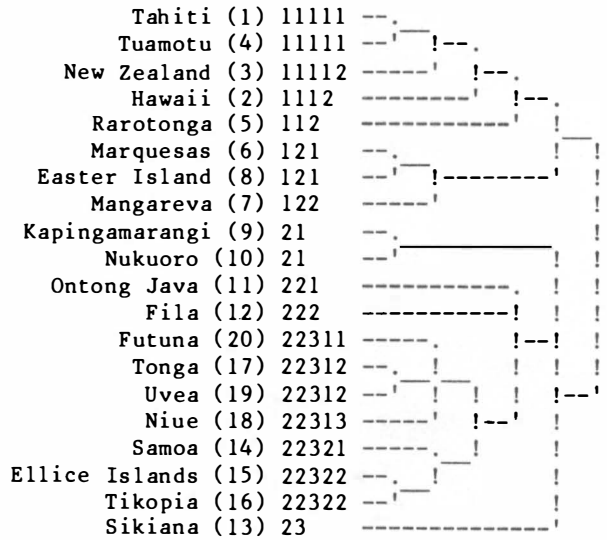
[7.4]

Reconstructed genealogy of the New Hebrides languages and dialects



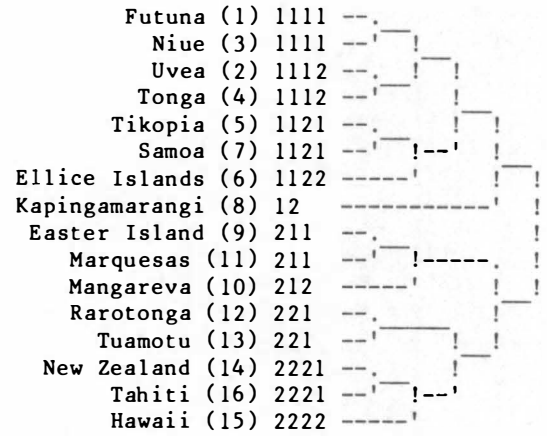
[8]

Tentative reconstruction of prehistoric migrations
in the New Hebrides



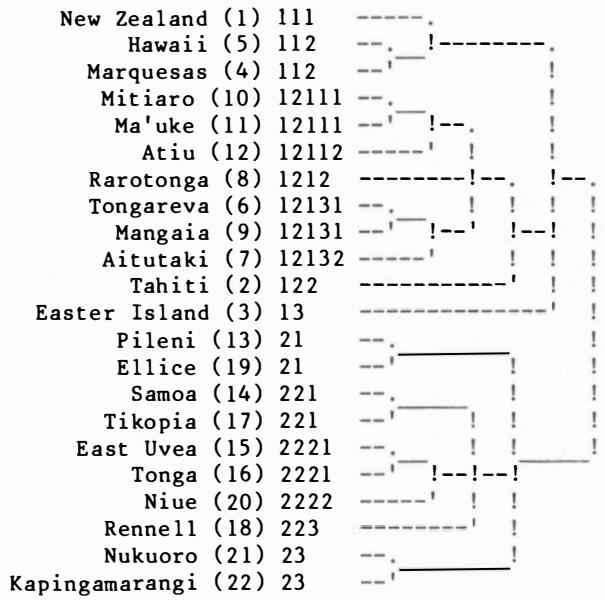
[9]

Reconstructed genealogy of 20 Polynesian languages
(data from Elbert 1953)



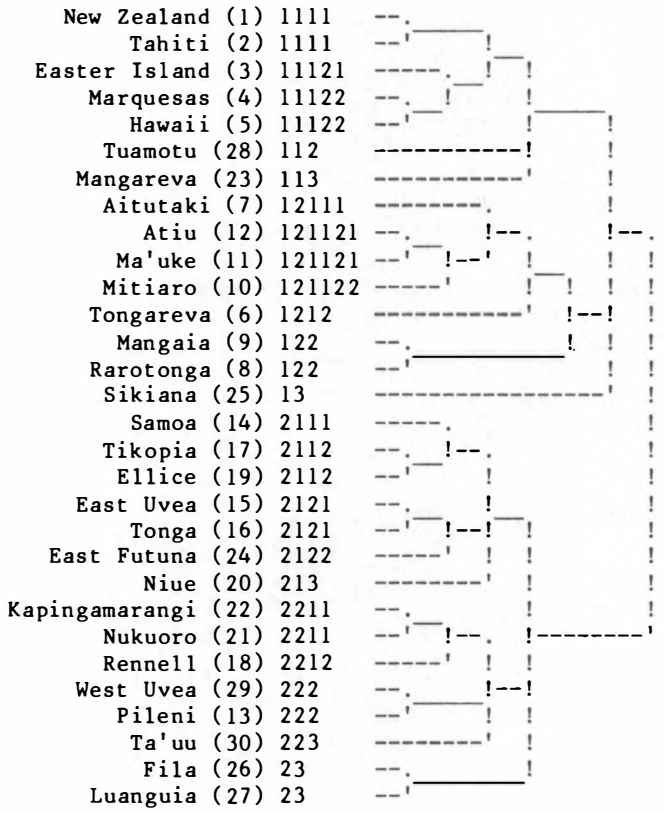
[10]

Reconstructed genealogy of 16 Polynesian languages
(data from Milke 1965)



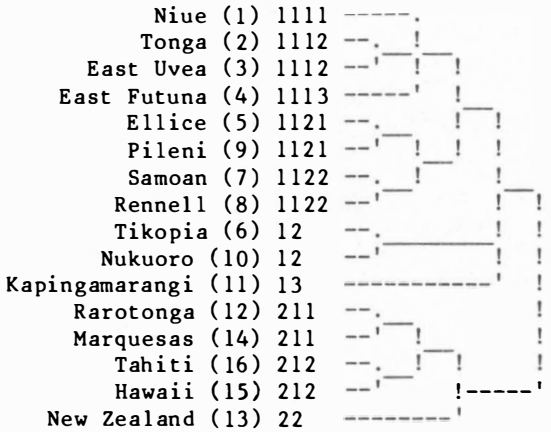
[11]

Reconstructed genealogy of 22 Polynesian languages
(data from Dyen, in Kirk and Epling 1972)



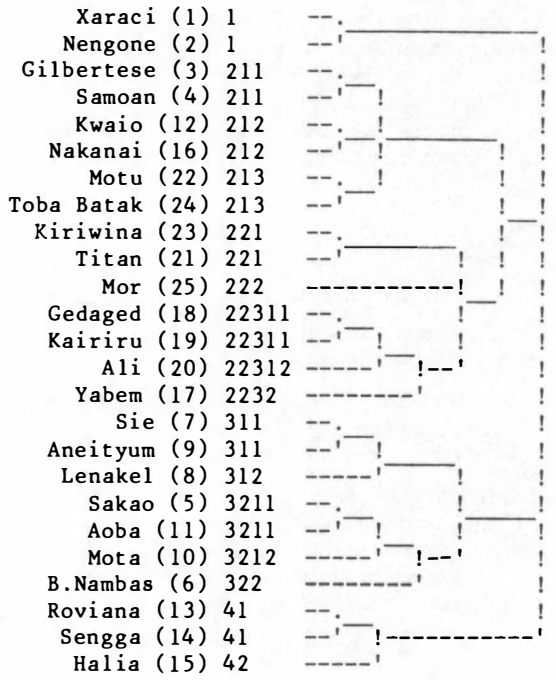
[12]

Reconstructed genealogy of 30 Polynesian languages
(data from Dyen, Elbert, and Emory in Kirk and Epling 1972)



[13]

Reconstructed genealogy of 16 Polynesian languages
(data from Biggs 1978)



[14]

Reconstructed genealogy of 25 Austronesian languages
(data from Tryon 1978)

APPENDIX

Comment by Dr. R.A. Blust, Leiden University:

The conclusion that Tanna is the area of greatest linguistic diversity in the southern New Hebrides is inconsistent with the hierarchy of splits proposed on p.293. If South New Hebridean divides into four primary branches of which two are found on Erromanga and only one on Tanna, then the area of greatest diversity (and hence the probable centre of dispersal) of the South New Hebridean group is not Tanna, but Erromanga. This conclusion is in no way affected by the fact that there may be more distinct languages on Tanna.

Reply to Dr. R.A. Blust:

As far as I can recollect, the reasons which led me to posit a centre of diffusion on Tanna rather than Erromanga were as follow:

1. There is a large number of reasonably diversified communalects on Tanna, only two on Erromanga. Evidence for a centre of dispersal on Erromanga thus seemed slim, resting on the survival of a single Erromangan language.

2. The lengths of the branches of the phylogenetic trees obtained by the methods outlined in the paper are not proportional to time depths: the trees are topological, not geometric, figures. Experiments on computer-produced language families (Guy 1980), had shown that, in accordance with the properties of topological figures, branches could be spurious (i.e. connect two nodes not separated in time). The branch connecting the Tannese node to the Sie-Aneityum-Ura node could thus be spurious.

However, the following possible counter-arguments have since occurred to me:

3. The dearth of present-day Erromangan communalects is due to extensive extinction, not lack of past diversification. Therefore, had most of them survived, Erromanga would probably have shown more than two languages as distant from one another as Sie and Ura are from each other now.

4. The other two centres of diffusion (South Santo and South Malekula) suggest a settlement pattern whereby the largest islands were populated first — perhaps because most likely to be spotted first. Erromanga also happens to be larger than either Tanna or Aneityum.

My choice of Tanna for a diffusion centre thus stemmed from a reluctance to fully trust my own methods: at the time the paper was written, I was aware of their many weaknesses, as I hope was made amply clear in the paragraphs entitled "A wider viewpoint and some critiques", and of what folly it would be to expect their blind application to yield final answers.

Dr. Blust's objection, however, is based on impeccable formal grounds and, after careful consideration, and taking into account counter-arguments 3 and 4 above, it no longer seems that points 1 and 2 are strong enough to warrant the exception I made. Let me, therefore, stand corrected.

Objection: by R.A. Blust

Page 287/288: Your discussion on p.287 suggests that graph [2] describes a theoretical model rather than an attested situation. But if this is true why should there be any scatter at all? As it is defined the linear equation is a function of differences in retention rate. If this is *all* that is being graphed in [2] then the scatter is unexplained. I gather that the scatter is in fact intended to represent the play of 'uncontrolled variables' (as intimate borrowing) which contribute to the results obtained in the real world. But then [2] is not graphing the theoretical model, but some approximation of what one might actually expect to find. Is it possible that you have conflated two different types of representation here?

Answer:

A function of differences in retention rates is *all* that is being graphed in [2], and the scatter does *not* represent the play of uncontrolled variables such as intimate borrowing!

If languages X and Y have, since the time they split, retained, say, 60% and 20% of their inherited vocabulary as represented by a sample wordlist, then the *most probable* amount of cognates shared between them is 0.6 x 0.2 = 0.12 = 12%. But this figure, although the most probable one, is in fact very unlikely to be the one observed: X and Y might have as many as 20% cognates and as few as none at all (true cognates, not apparent cognates due to intimate borrowing). Toss a fair coin 100 times. The most probable number of heads is 50, but it will be a rare occurrence indeed when you get *exactly* 50 heads. The scatter in [2] is meant to express this fact: if X and Y have retained 60% and 20% of their inherited vocabulary, they may share between 0% and 20% cognates, but, if you are forced to take a guess at the exact proportion that they indeed share, then your best bet is 12% (but you'll need very long odds indeed to make this a fair betting proposition).

Objection: by R.A. Blust

Page 294: You take the results illustrated in diagram [14] as commending the procedure adopted to achieve them. The cases you choose as supporting the validity of the method may seem intuitively probable, but there are at least two two-language subgroups in diagram [14] which are intuitively improbable: 1) Motu + Toba Batak, 2) Kiriwina + Titan. One of these (Motu + TB) is demonstrably in error, as it is contra-indicated by substantial phonological, lexical and grammatical evidence that Motu belongs with *all* other languages in diagram [14] apart from TB and Mor (the only non-Oceanic languages in the sample). If the Eastern Malayo-Polynesian hypothesis is correct the Oceanic languages + Mor further subgroup together as against Toba Batak. Error is more difficult to demonstrate in the second case, but I know of no qualitative evidence that Kiriwina and Titan subgroup together. When I wrote *The Proto-Oceanic palatals* I made a fairly serious effort to determine the position of the Admiralty subgroup within Oceanic, and failed to uncover support for any larger grouping short of Oceanic itself. Your results in the Polynesian cases similarly fail to agree with the generally accepted qualitative arguments advanced by Pawley (1966) that the first split within PN is between Tongic (Tongan, Niue) and Nuclear Polynesian (the rest).

Reply:

You are right. You must remember that my paper says that results obtained from a small number of languages should be viewed with extreme suspicion. The only Polynesian data set of resonable size (30 lists, Kirk and Epling 1972) is far from reliable as it is a mixture of straight cognate counts, averages, and results of some mathematical transformation.

As for the Austronesian set, it is too small (25 lists) for such a motley sample, and is unbalanced (e.g. only one representant of Polynesian, but seven of New Hebridean).

Since writing this paper (May-June 1980), I have discovered that a result of the properties of linear-correlation co-efficients is that an isolated language is likely to be randomly misclassified with some other language. Thus for instance the two non-Oceanic languages in the sample of 25 Austronesian languages used are likely to be wrongly grouped within some Oceanic subgroups.

BIBLIOGRAPHY

BERGSLAND, Knut and Hans VOGT

1962 On the validity of glottochronology. *Current Anthropology* 3/2:115-153.

BIGGS, Bruce

1978 The history of Polynesian phonology. In: S.A. Wurm and Lois Carrington, eds *Second International Conference on Austronesian Linguistics: proceedings*. *PL*, C-61:691-716.

DYEN, Isidore, A.T. JAMES and J.W.L. COLE

1967 Language divergence and estimated word retention rate. *Language* 43/1:150-171.

ELBERT, S.H.

1953 The internal relationships of Polynesian languages and dialects. *Southwestern Journal of Anthropology* 9/2:147-173.

GUY, Jacques B.M.

1978 Proto-North New Hebridean reconstructions. In: Wurm and Carrington, eds 1978:781-850.

1980 *Experimental glottochronology: basic methods and results*. *PL*, B-75.

HARTIGAN, John A.

1974 *Clustering algorithms*. New York: Wiley.

KIRK, Jerome and P.J. EPLING

1972 *The dispersal of the Polynesian dialects: explorations in phylogenetic inference from the analysis of taxonomy*. Working Papers in Methodology 6. Chapel Hill: University of South Carolina.

MILKE, Wilhelm

1965 Experiments in matrix reduction: applied to Austronesian data. *Lingua* 14:443-455.

SWADESH, Morris

1950 Salish internal relationships. *International Journal of American Linguistics* 16:157-167.

TRYON, D.T.

1976 *New Hebrides languages: an internal classification*. *PL*, C-50.

1978 The languages of the New Hebrides: internal and external relationships. In: Wurm and Carrington, eds 1978:877-902.

Guy, J.B.M. "Bases for new methods in glottochronology". In Halim, A., Carrington, L. and Wurm, S.A. editors, *Papers from the Third International Conference on Austronesian Linguistics, Vol. 1: Currents in Oceanic*.
C-74:283-314. Pacific Linguistics, The Australian National University, 1982. DOI:*10.15144/PL-C74.283*