Veröffentlichungen des Seminars für Indonesische und Südseesprachen der Universität Hamburg

Band 17

Sonderdruck

GAVA'

Studies in Austronesian Languages and Cultures
Studien zu austronesischen Sprachen und Kulturen

dedicated to
HANS KÄHLER
gewidmet

edited by/herausgegeben von
Rainer Carle · Martina Heinschke · Peter W. Pink
Christel Rost · Karen Stadtlander

1982

DIETRICH REIMER VERLAG · BERLIN

MICRO- AND MACRO-SUBGROUPING: CRITERIA, PROBLEMS, AND PROCEDURES

R. David Zorc, *Darwin*

Hans Kähler's conspectus of "Austronesian Comparative Linguistics and Reconstruction of Earlier Forms of the Languages" (1978) offers a good basic outline of procedures and caveats in attributing etyma to PAN or its subgroups. Almost three decades ago, Dyen (1953, p. 580) discussed "[t]hree subgrouping procedures ... now available to the comparatist: (1) judgment by inspection; (2) discovery of exclusively shared (non-accidental) innovations; (3) lexicostatistic dating". In the last decade a number of subgroupings of various An languages have emerged, from the PAN level (Dahl 1976; Blust 1977; Harvey 1979) through to various low-level subgroups, such as North Sarawak (Blust 1974), Bikol (McFarland 1974), Bisayan (Zorc 1977), Mansakan or Proto-South-East Mindanao (Gallman 1979), Danaw (Allison 1979), Minahasan (Sneddon 1978), Tsouic (Tsuchida 1976), Malayo-Javanic (Nothofer 1975), Central Cordilleran (Reid 1974), Mangyan (Zorc 1974a) to cite only a few. Several methods have been applied in arriving at a subgrouping hypothesis, some of which are newly devised, and deserve summary and comment.

If one imagines an ideal Austronesian family tree (one that we are striving for, but are yet to achieve) micro-subgrouping involves looking downwards from above, while macro-subgrouping involves looking upwards from below. More explicitly, the term *micro-subgrouping* here refers to determining the immediate relationship of genetically-close languages, while *macro-subgrouping* refers to determining the relationship of micro-subgroups to one another.

1. JUDGMENT BY INSPECTION

Any linguist involved in research on two or more languages cannot help but note resemblances and make comparative judgments of a typological or genetic nature. However, even the most rudimentary subgrouping procedure must take account of *shared differences* (not similarities) as basic criteria. If two or more speech varieties consistently share differences in the areas of phonology, lexicon, and grammar from other languages, then the multiplicity of shared differences seems to be great enough to justify subgrouping them together. Such subgrouping by inspection "depends for its validity on the probability that some of the exclusively shared differences are not due to independent linguistic change, separate borrowing, or retention" (Dyen 1953, p. 582). Unfortunately, authors who speak only of shared similarities do not present any special case for subgrouping, e.g., for Mangyan (Barbian 1977) or the Bisaya of Borneo and the Philippines (Araneta and Bernad 1960). This phenomenon of shared differences has also been described in terms such as "contrasting evidence" (Zorc 1974a, pp. 59f. in excluding Mangyan languages from the Central Philippine subgroup) or "contrastive evidence" (Zorc 1978, p. 520 in excluding Kamayo from Bisayan, but including it in Mansakan).

The Danao, Manobo, Subanon, and Mansakan languages of Mindanao together reflect an etymon *si+dan 'they', which differs from PMP *si+Da in having an additional *-n. These same languages also reflect an etymon *ʔəta:w 'person', which differs from PAN *Ca:ɸu[H] in having an innovative prefix. One discovers that they also share an etymon *ʔəbəl 'smoke', but cognates are found in Tboli *kəbəl*, Kanakanabu *ʔəəvə*, Rukai-Budai *əbə́lə*, so that this represents a shared retention of PHF *qəbəl. In judgment by inspection, a hypothesis takes shape on the basis of two qualitative differences, but is slightly shaken by the discovery of a shared retention. Can a macro-subgroup comprising Danao, Manobo, Subanon, and Mansakan be established? If the two features are innovations, and not loans, such a hypothesis could be put forward. But other tests and counter-hypotheses become critical. For example, Zorc (1977) has shown that Mansakan shares a number of innovations of considerable quality with Bisayan, Bikol, and Tagalog, so that it forms part of a Central Philippine subgroup. Gallman (1979) considers some of Zorc's South Bisayan members to be part of an immediate subgroup (Proto East Mindanao) along with Mansakan. Danao, Manobo, and Subanon do not share any of the innovations discussed in either study, nor do Bisayan, Bikol, or Tagalog reflect *si+dan, *ʔə+taw, or *qəbəl (since they retain *sĭda 'they', *ta:ɸuh 'person', and *qăsu(h) 'smoke'). What then of the position of Mansakan? With which macro-subgroup does it belong? Judgment by inspection shows that Mansakan phonology, grammar, and lexicon is of the Central Philippine rather than of the Danao, Manobo, or Subanon type, so that its startling agreement with the latter languages is enigmatic. Finer tools are therefore necessary to establish the immediate lines of its genetic affiliations.

2. LEXICOSTATISTICAL CLASSIFICATION

Swadesh (1952) pioneered the methods of lexicostatistics and glottochronology by means of a 200-meaning list. Swadesh (1955) refined the method, and introduced a 100-meaning list. The procedures have been discussed in great detail (Hymes 1960), attacked (Bergsland and Vogt 1962, Teeter 1963), and defended (Dyen 1962, 1963, 1964). Scholars have noted the variability of retention rates (Dyen, James, and Cole 1967 for Austronesian), and the possibility of gross error due to conservative vs innovative language groups (Blust 1981). While this method counts the sum of both retentions and innovations without distinguishing between them, it is at least a test of the synchronic if not the purported diachronic inter-relationships of languages.

A number of lexicostatistical studies have been done on Philippine languages alone (Dyen 1953, Thomas and Healey 1962, Dyen 1965, pp. 29-33, Walton 1977). The results do not always agree because of approaches to borrowing, interpretations of highest, average, or lowest undistorted percentages, data gathered from various sources, lacunae in the lists, and the type of list used (Swadesh 200, Swadesh 100, or SIL 372). All of these studies reflect errors rectifiable by recourse to different statistical approaches (e.g., accounting for borrowing by inflated percentages or innovation by deflated percentages) or, more importantly, to items on the list which indicate innovations, and therefore genuine lines of genetic relationship, such as Manobo

elements in Kagayanen (Zorc 1974b, pp. 414-18). Sneddon (1970) has made an excellent study of Minahasan languages on this basis, that is worthy of emulation since the data and cognate decisions are presented along with a thorough discussion of procedures and problems.

Lexicostatistics can be an important and valid *first step* in making a subgrouping hypothesis, so long as the following caveats are observed.

(1) Care is needed in gathering the data. When working with someone else's data, it is difficult to determine the exact meaning of a form, and whether it is correctly matched with the other forms being compared. When gathering one's own data, learned informants can give cognates or parrot prestigious source language forms. Speech registers and language levels must also be considered (e.g., Javanese Krama versus Ngoko).

(2) Care is needed if revising a list becomes necessary. Swadesh (1955) worked out a 100-meaning list "in the realization that quality is at least as important as quantity", although the new list was discovered to have a higher retention rate (Swadesh 1955, p. 127; see also Zorc 1977, pp. 174ff.). The introduction of longer, more culturally-oriented lists (Walton 1977, Barbian 1977) does not account for a higher probability of borrowing and indeterminate retention rates. The substitution of other lists (e.g., body parts or kin terms) ceases to be lexicostatistics as anyone knows it, but may have value as a contrastive or supplementary method.

(3) Care is needed in scoring, since one must be aware of standard sound changes. Just because forms are too similar does not mean they are non-cognate (see Witucki 1974), especially where conservative phonemes are reflected identically. Furthermore, partial similarity is not sufficient for a positive score, e.g., Tag *dugó?* 'blood' < PAN *ZŭRuq 'sap' while Chamorro haga? < PAN *Da:Raq 'blood' (Witucki 1974, p. 67). Zorc (1977, p. 174) introduced a principle of morphological identity, whereby words had to be *cognate in toto* such that frozen morphological formatives or irregular sound shifts led to negative scores if not shared by language pairs (see Table 2, note).

(4) Borrowings should be accounted for systematically and with care. If they are clearly identifiable, they can be scored "ø" (neither cognate nor noncognate), but this reduces the number of usable items in the list. Or they can be counted as cognate, and adjustments can be made by compensation for any skewed or inflated scores in evaluating the resulting tables. It should be noted that borrowing is one form of innovation and a legitimate (albeit problematic) element of language change.

(5) Differences in retention rates need to be considered, either with regard to individual items in the list, or to a given language's overall scores. Blust (1981, p. 11) noted: "... major features of the structure of Dyen's classification of the Austronesian languages can be seen to follow in a straightforward manner from significant variation in retention rate rather than from real differences in separation times." However, one surprising result of his study "is the discovery that most Oceanic languages probably do not show unusually low retention rates - rather, the languages of the Philippines and western Indonesia (with some notable exceptions) show unusually high retention rates". (Blust 1981, p. 56.)

The Mangyan languages of Mindoro provide a stimulating case for micro- and macro-subgrouping. Zorc (1974a) proposed two groups: North Mangyan (Iraya, Alangan, Tadyawan/Balaban) and South Mangyan (Hanunoo, Buhid/Buid). It was not possible to demonstrate convincingly that these two groups could be macro-subgrouped. Data from Taubuid/Batangan were not then available, but Pennoyer (1976) has since shown that Taubuid subgroups closely with Buhid. He is sceptical that this micro-subgroup belongs in an immediate macro-subgroup with Hanunoo. Using a modified version of the Swadesh 100-meaning list (Table 1), I have expanded my previous study to include Taubuid (Table 2). There is no doubt that Buhid and Taubuid are closely related; but the next highest scores are with Hanunoo, not with Tadyawan, Taubuid's immediate neighbour to the north. Furthermore, the scores for Buhid and Taubuid parallel each other in relationship to all other Mangyan languages. Tadyawan, Alangan, and Iraya exhibit a chaining relationship, but the Iraya-Tadyawan score is exceptionally low, and therefore problematic. The two basic Mangyan groups still

TABLE 1. THE SWADESH 100-MEANING LIST (MODIFIED FOR MANGYAN LANGUAGES)

all	*fear	male/man	sit	OMITTED FROM SWADESH (1955)
ashes	feather	many	skin	bark = skin
belly	fingernail	meat/flesh	sleep	claw → fingernail
big	fire	moon	small	lie = sleep
bird	fish	mountain	smoke	horn
bite	to fly	mouth	stand	seed
black	foot	name	star	yellow
blood	full	neck	stone	
body	give	new	sun	
bone	good	night	swim	
breast	green/raw	nose	tail	
burn	hair (head)	not (future)	this	ADDED SINCE ZORC (1977)
cloud(y)	hand	one	that/you	fear
cold	head	*palm (hand)	thou	palm (hand)
come/arrive	hear	person	tongue	right (side)
die	heart	rain	tooth	shoulder
dog	hot/warm	red	tree/wood	
drink	I	*right (side)	two	
dry	kill	road/trail	walk	
ear	knee	root	water	
earth	know (how)	round	we (exclusive)	
eat	leaf	sand	what?	
egg	liver	say	white	
eye	long	see	who?	
fat/grease	louse	*shoulder	woman	

Note. A discussion of special applications of the Swadesh list is found in Zorc (1977, pp. 171-73). Four items have replaced that previous list. Mangyan languages either have a loan (Tag *dilâw*) or an identical form as for 'red', so 'yellow' was not considered diagnostic. Several languages do not have different forms for 'lie-down' and 'sleep'. 'Seed' appears to have been replaced as a culture item (since *h-losing dialects have Tag *binhî?*). 'Horn' is likely to have been an introduced concept, with all languages reflecting *suŋáy. Each of the four replacements have a diagnostic bias: NMg *limu, SMg *dalá? 'fear', PMG *dalu:kap 'palm (of hand)', NMg *pamalaŋ, SMg *sikə̂n 'right (side)', SMg *la:bay, PPH qa:baRah 'shoulder'. Despite these revisions, scores remain quite similar to the comparison cited in Zorc (1974a, p. 562) and indicate the same subgrouping proposed therein, and in Dyen (1965, p. 30).

TABLE 2. LEXICOSTATISTICAL SCORES: MANGYAN LANGUAGES PLUS DATAGNON, TAGALOG, ILOKANO, AND KAPAMPANGAN

Hanunoo									
58^{-1}	Buhid								
51	70	Taubuid							
43^{-3}	41^{-0}	42	Tadyawan (Balaban)						
47^{-0}	41^{-2}	40	63^{-0}	Alangan					
41^{-0}	36^{-1}	38	43^{-4}	65^{-1}	Iraya				
49^{-5}	34^{-6}	30	34^{-7}	39^{-5}	45^{-3}	Datagnon *(equated with Aklanon)*			
47^{-7}	37^{-9}	35	37^{-7}	44^{-6}	48^{-3}	[61]	Tagalog		
38^{-6}	32^{-6}	[22]	30^{-11}	32^{-8}	33^{-8}	35^{-9}	[49] $^{-3}$	Kapampangan	
43	37	33	34	34	32	35	38	35	Ilokano

Note. Scores from this comparison are slightly lower than those reported in Zorc (1974a, p. 583) because of the introduction of the principle of morphological identity (see Zorc 1977, p. 174), and not because of the changes introduced in the list (see Table 1). By this principle, Aln, Tdy *maknu?* (with dissimilation) are scored negatively when compared with Iry *kapnu?*, Tau *apnu*, Buh *napnu* (with syncope), and both of these with Han, Dtg *punú?*, Tag *punô?*, all of which derive from PAN *pə̆nuq 'full'. Similarly, the addition of the prefix in Buh *mwayan*, Tau *moyan* leads to a negative score when compared with Aln, Iry *?udan*, Han, Dtg *?urán*, Tag *?ulán* 'rain' even though all ultimately derive from PAN *qŭZaN 'rain'.

stand: South Mangyan includes Hanunoo and Buhid-Taubuid; North Mangyan remains unchanged. I have included scores from Datagnon (a West Bisayan language neighboring Hanunoo), Tagalog (a Central Philippine language which has influenced all Mangyan languages as a donor), Kapampangan (a South Luzon language), and Ilokano (a north Cordilleran language) to highlight the difficulties in proposing a macro-Mangyan subgroup, at least on a lexicostatistical basis. Inter-influence and borrowing has obviously blurred the picture throughout Mindoro; but Hanunoo appears to be a highly retentive language (note the high score with Ilokano and generally high scores with the other languages presented here and in Zorc (1974a, p. 583)); and Taubuid appears to be a highly innovative language (note the low score with Kapampangan and generally lower scores with most other languages). Nevertheless, such evidence from lexicostatistics offers a good starting point for further genetic subgrouping hypotheses. Even when the scores appear to go wrong, the results are of at least sociolinguistic and, in this regard, historical importance, i.e., they indicate non- or post-genetic influences.

3. DISCOVERY OF EXCLUSIVELY SHARED (NON-ACCIDENTAL) INNOVATIONS

According to the strictist tenets of the comparative method, genetic relationships can only be based on the qualitative evidence of shared innovations wherein factors such as chance, borrowing, or convergence have been (totally) eliminated. When such qualitative innovations have been isolated, they should be of sufficient number as to be convincing; it is in this regard that quantity counts, although scholars are not at all in agreement on how much is enough. In general, the better the quality, the less need for quantity; hence, a heavier burden is put on the scholar to evaluate and rank the significance of the innovations he discovers than to produce a massive list of them.

Phonological innovations, particularly sound shifts, are of generally poor quality, because they can happen independently, and are also subject to influence even across genetically-wide linguistic boundaries (witness the loss of syllable-final *r* in Singaporean Bahasa Melayu after prolonged contact with British English). Zorc (1977, pp. 219-21) and McFarland (1974, pp. 78-83)

discuss the weaknesses of subgrouping Central Philippine languages by phonological criteria alone, but both note that in conjunction with other criteria they can be of some use. Of greater use are complex phonological innovations, e.g., metathesis of *lC, *hC, and *ʔC clusters in Bisayan (Zorc 1977, pp. 241ff.) or the assimilation of *ld clusters to *ll in Mansakan (Gallman 1979, pp. 10, 21f.). Pennoyer (1976) shows that Buhid and Taubuid share four shifts: *p > f, *k > h (thence > ø in Tau), *h and *ʔ > ø, *D and *j > y, some of which have happened independently in Philippine languages, but not all in the same language group; but the complex innovation he discusses is the restructuring of *-ai- and *-au- sequences, e.g., *ma-ʔuran > Buh *mwayan*, Tau *moyan* 'rain', *ma-ʔinit > Buh *myanit*, Tau *menit* 'sun', *kaʔun > Buh *kwan*, Tau *kon* 'eat'. Although it is clear that Taubuid underwent a subsequent change (*wa > o, *ya > e), the metathesis of the original sequence, as evidenced by Buhid, represents a complex phonological innovation of considerable weight. Blust's discussion (1969, 1973, 1974, 1980) of vowel deletion among languages of North Sarawak characterizes a complex and qualitative innovation, but projects certain anomalies into the proto language of highest order (PAN) which are in need of external substantiation, e.g., PNS *təbSu < PAN *təbuSu or *tə̆buS 'sugarcane', PNS *bSaq 'water' < PAN *baSaq or PMP *bă̆haq 'flood', PNS *[dj]Sən < PAN *[dzZ]əSə[nN] or PHN *də̆qən 'downward pressure', PNS *pədSu < PAN *qapəjuS(u) or PAN *qa(m)pə̆juø 'gall(-bladder)', PNS *idSuŋ < PAN *ijuSuŋ or PAN *qĭjuŋ 'nose', PNS *bSaq < PAN *(ba)baSaq or PMP *baqbaq 'mouth', PNS *bSaR < PAN *baSaR or PHN *bă̆haR 'loincloth', etc.

Lexical innovations are difficult to evaluate. It is practically impossible to distinguish a common from a spread innovation, and, in the case of conservative phonemes, to isolate a borrowing. Furthermore, any given form may be a retention lost everywhere else or as yet undiscovered in another language. However, certain precautionary measures may be taken to insure both care and quality (see Zorc 1977, pp. 234f.):

(1) Limit forms to basic vocabulary and avoid items of trade or culture that could freely pass from one language to another.

(2) Dismiss forms with phonological irregularities, i.e., not in conformity with the standard reflexes worked out for a given language, e.g., *h* in a language that loses *h, *r* in a language where *R > *g* or *R > *y*, etc.

(3) Reconstruct, wherever possible, an etymon for a given meaning at the earliest possible stage, e.g., 'blood' was PAN *Da:Raq, so SPh *dŭgu? or Manobo *lăŋəsa are innovations.

(4) Consider the character and quality of each lexical innovation, including its geographical and linguistic distribution, potential spread, etc.

(5) Determine if the innovation is a formal or semantic one, i.e., an old form has changed meaning (PAN *bărəq 'abscess' > SBs *băga? 'thick', PAN *ba:RaH 'embers' > Gubat *băga* 'red') or a new form has been coined from previously unknown material (PBS *hăbag 'abscess', SPh *pŭlah 'red'), and if the change could happen independently (note Ivatan *ma-vaya?*, Ilokano *nala-ba:ga* 'red').

In this light, Pennoyer's conclusions are perhaps unduly pessimistic:

> It is tempting to use lexical innovations as supportive evidence for drawing immediate genetic connections between Mindoro languages ... This type of comparison is interesting, but inconclusive due to a number of factors including the inadequacies of incomplete lexical lists from each language, borrowing, and loss and replacement. Thus, little weight should be placed on lexical 'innovations'. (1976)

The PAN word for 'star' is reconstructed as *bi(n)tu:q-ən. It is an item of basic vocabulary, and the form is reasonably retentive (Dyen, James and Cole 1967, #67 out of 196 ranked items). In all known Mangyan languages it has been replaced: Hanunoo *paŋgasán*: Buhid *faŋgasán*, Taubuid *galeme* : Tadyawan *galaymay*, Alangan, Iraya *magirêm*. Lexicostatistical and other evidence (of varying quality, consult Zorc 1974a) would agree that the Han-Buh form is a South Mangyan innovation, and that the Aln-Iry form is a North Mangyan innovation. There is no doubt that the Tau and Tdy forms are cognate and represent an innovation, but there is:

(a) no (other?) qualitative evidence that Tau and Tdy belong in an immediate subgroup, (b) no lexicostatistical evidence that they share any especially close genetic connection, (c) the fact that the two languages border one another, and (d) the observation that [g] is problematic in Tadyawan since *R > *y* in most basic vocabulary, while *R > *g* in Taubuid. Therefore, the North and South Mangyan innovations stand along with other evidence, but it is clear that a Tadyawan innovation (*galaymay) was borrowed early into Taubuid, and then underwent the *ay > *e* sound shift. Having followed all five precautionary measures (above), this analysis seems reasonably sound and straightforward.

Perhaps the most qualitative evidence is found in core grammatical items. One kind is innovative morphological formatives, e.g., *ma- in the Buhid and Taubuid words for 'rain' and 'sun' (Pennoyer, op.cit., above). Others include innovations amongst pronouns, e.g., WBs *ta:na 'he/she' replacing PMP *s+ĬØa (Zorc 1972, p. 118), or deictics, e.g., South Cordilleran *tan 'second person' and *man 'third person' (Zorc 1978, pp. 511-13), which would then compel the subgrouping of Ilongot with South Cordilleran, at least at a macro-level (Zorc 1978, Reid 1979, contrast McFarland 1980). An important facet of grammatical change is that it is systematic, such that if a new base is coined, it usually pervades the entire paradigm; hence, while one speech variety may borrow a member of such a paradigm, it would be most unlikely to borrow the entire system. However, it is crucial that the scholar identifies the original and then the innovated system.

4. OTHER METHODS DEVISED OR EMPLOYED MORE RECENTLY

The use of functors as a means of subgrouping languages has been applied by McFarland (1974), Zorc (1977, 1978), and Allison (1979). Function words form the core of any given speech variety, and identify it as a specific language; they are also of very high text frequency. Therefore, a means of comparing languages on the basis of their grammatical systems should prove invaluable in comparative work, although some scholars have expressed scepticism (Pallesen 1977, Gallman 1979). Of particular interest is the phenomenon where languages compared by functor analysis yield a higher score than by lexicostatistics (cases thus far