Subgrouping

Prepublication draft. To appear in:
The Wiley Blackwell Companion to Diachronic Linguistics
Adam Ledgeway, Alexandra Simonenko, Anne Breitbarth, Edith Aldridge, Joseph Salmons, Katalin É. Kiss (eds.)

Daniel Kaufman
Queens College, CUNY & ELA
dkaufman@qc.cuny.edu

Abstract
This chapter gives an overview of subgrouping within the comparative method with a focus on several successes within the Malayo-Polynesian subgroup of the Austronesian family, contrasting successes which crucially rely on innovative phonological and morphological change with a controversial subgroup that is adduced purely through putative lexical innovations. A further comparison to subgrouping approaches rooted solely in shared cognacy rather than linguistic innovation show that they merely recapitulate well-known contact relations rather than uncovering ancient phylogenies.

Keywords
[comparative method, stammbaum, language classification, Austronesian, Malayo-Polynesian, computational phylogeny]

1. Introduction
This chapter investigates the internal structure of language families and how decisions are made to include or exclude languages under a particular node in a family tree, focusing on subgrouping via the comparative method with only brief remarks on other methods. For an introduction to Bayesian approaches not covered here, see COMPUTATIONAL PHYLOGENETIC MODELS as well as the discussion in THE FAMILY TREE MODEL.

Our first order of business should be to clarify what subgroups represent. Greenberg (1957) asks, "given the relationship of A, B, and C, is the distance between A and B equal to, or less than, the distance from A to C?" If A and B are "closer", Greenberg explains, it signals that A and B can be traced to a social community that branched off and underwent certain linguistic changes, which it then bequeathed to its descendants. But the relation between a linguistic subgrouping and the branching out typically associated with human migration need not be one to one, as not all social splits entail linguistic innovations that can be identified with certainty. Conversely, linguistic innovations can take place without migration, that is, in speech varieties other than (geographic) dialects, such as sociolects, ethnolects, religiolects and genderlects. Family trees often shed a crucial light on population history and migration but it must be kept in mind that a subgrouping hypothesis only attempts to explain a subset of linguistic facts, those whose directionality and inheritance are relatively clear and compatible with "clean speciation". It is thus a partial reflection of social splits which can then be coordinated with evidence from other areas to build a holistic picture of population history and movement.

The literature on subgrouping, with its origins in the study of European languages, primarily Indo-European, is both old and vast. Here, we restrict ourselves to the Malayo-Polynesian (MP) branch of the Austronesian family, which has been at the center of several lively methodological and theoretical debates that may be less familiar to the generalist. The Austronesian languages comprises a family of over 1,200 languages that span half the globe, from Madagascar in the west, to Easter Island in the east. The historical development of Austronesian languages has been worked out in extremely fine

detail beginning with Dempwolff (1934). Today, Proto-Austronesian easily stands as one of the most firmly and broadly reconstructed languages of all. A history and summary of the scholarship can be found in Blust (2013).

The chapter introduces the logic of subgrouping via the comparative method in §2 and moves on to a series of instructive case studies in §3. We first examine several uncontroversial successes in subgrouping, where the comparative method has led us to surprising conclusions that would have been otherwise unreachable and then proceed to a controversial case that pushes lexical evidence to its limits. Finally, §4 offers a critical review of similarity-based approaches and §5 concludes with future directions.

2. Subgrouping and the comparative method

The primary goal of the comparative method (see Bostoen, this volume), is the identification of languages related by common descent and the reconstruction of their common ancestor. To a large extent, traditional subgrouping is simply the result of applying the comparative method recursively within a single family of related languages to yield an articulated family tree. Karl Brugmann (1884: 253) is widely recognized as the first to formulate subgrouping as a three stage process:

Stage 1: Identifying similiarities
Stage 2: Identifying exclusively shared similarities
Stage 3: Identifying exclusively shared innovations

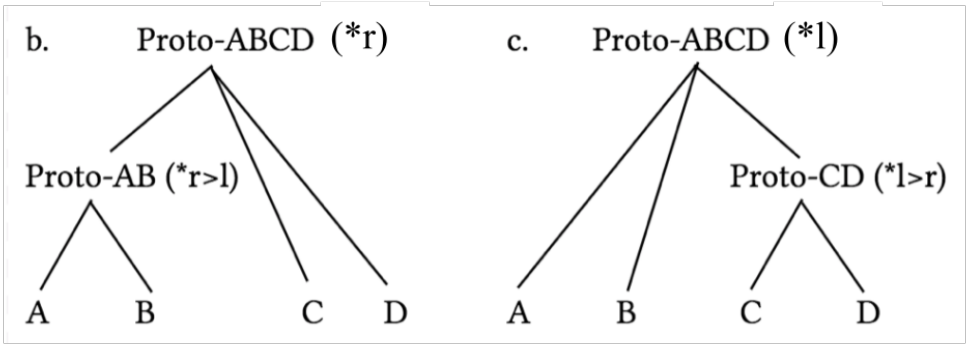
It is only evidence from Stage 3 of this process that provides positive evidence for subgrouping a set of languages together. Mere similarities, even exclusively shared ones, do not necessarily indicate a special relationship between a set of languages. Similarities may be inherited from an ancestor by a much wider group of languages and subsequently lost by various members. While the loss may indicate a shared development, the retention simply continues to reflect the initial state. On the other hand, exclusively shared innovations between a set of languages suggests descent from a unique ancestor which underwent those changes before breaking up into distinct language communities. However, not all diachronic changes are equally probative. For instance, a historical change involving intervocalic voicing of stops is highly natural, and therefore common, but a change such as *w > c is highly idiosyncratic, without any clear grounding in articulation or perception. The more idiosyncratic a change is, the more valuable it is for subgrouping because the chances of it occurring multiple times independently are far smaller. Common changes, on the other hand, only provide weak evidence for subgrouping because such changes are just as likely to arise through parallel but independent developments. Researcher bias and general subjectivity in determining the probative value of a change remains a standing challenge to using the comparative method for subgrouping.

A priori assumptions about which languages *should* be conservative based on external factors have turned out to be mostly incorrect throughout the history of linguistics. Among the Malayo-Polynesian languages, Kawi (Old Javanese) manuscripts and inscriptions had originally been assumed to represent an ancestral Austronesian language that gave rise to much of the modern diversity across the region (von Humboldt 1836). This was based on the fact that Kawi was a revered literary language that had been preserved in writing for over a thousand years. However, Kawi was ultimately shown to be just as innovatory as many modern Austronesian languages and the timespan afforded by even the oldest inscriptions was not nearly sufficient to find anything resembling a common ancestor to the Malayo-Polynesian languages. Across many language families, the most historically conservative languages have often been found in the most unexpected places, hearkening back to Sapir's (1921: 219) dictum, "When it comes to linguistic form, Plato walks with the Macedonian swineherd…".

The key role of uniquely shared innovations in subgrouping is predicated on being able to distinguish innovations from retentions, but this is not a trivial task. Given the hypothetical data in (1), where languages A and B show a lateral corresponding to a trill in languages C and D, there are no a priori methods for knowing whether this supports grouping A and B together on the basis of the innovation *r > l or grouping C and D together on the basis of *l > r. These two possibilities are shown as stammbaum (family trees) in (1b) and (c), the most popular means of visualizing phylogenetic relations since Schleicher (1853).

(1)a.

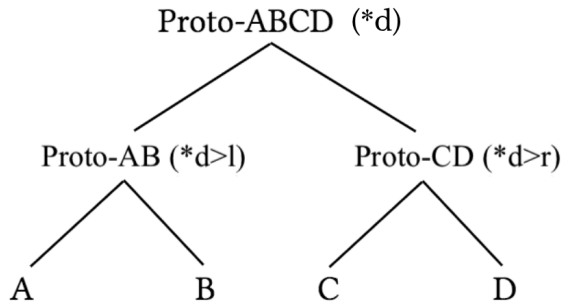
A	B	C	D
l	l	r	r



Fortunately, the directionality of change is not always so ambiguous. Comparative evidence shows that certain correspondences, such as [s] and [h], are far more likely to result from one change, *s>h, than from the other, *h>s. Such unidirectional or nearly unidirectional changes are clearly very useful in distinguishing innovations. Thus, given ten languages, only two of which show an *s* reflex corresponding to *h* in the other eight, the null hypothesis would still be to reconstruct *s and treat the innovation *s > h as evidence for a subgroup including the eight *h* languages. The *s* reflex would offer no support for subgrouping the two *s* languages together because this reflex is a retention. Such scenarios are considered to vitiate the notion of "majority wins" in reconstruction and subgrouping. Concretely, it is incorrect to suppose that a proto-phoneme should be reconstructed as the reflex that happens to be the most widespread.

The null hypothesis in reconstruction and subgrouping takes into account the principle of least changes, simply as a matter of Occam's razor, which disallows positing more historical changes than absolutely necessary. Without Occam's razor, retentions could easily be recast as innovations giving rise to unwarranted subgroups. For instance, in (2), using the same scenario as above, it is hypothesized that both intervocalic *l* and *r* come from a different source, *d, thus implying two innovation-defined subgroups, Proto-AB and Proto-CD. While such a history is possible, given the widespread nature of changes such as *d > l and *d > r, it is just one of many possibilities, none of which should be posited without supporting evidence.

(2)



Another key source of directionality in change is the well supported observation that phonological mergers are far more common than unconditioned splits. Where Language A reflects two or more reflexes corresponding to a single reflex in Language B and there is no evidence for a conditioned split in the former language, we have evidence for an innovatory merger in Language B. Such a situation is seen in the comparison in (3).

(3)	Malay	Cebuano		PMP/PWMP reconstruction
a.	dʒalan	**d**alan	'path'	*zalan
	dʒedʒal	**d**ul**d**ul	'to stuff'	*zelzel
	endʒak	un**d**ak	'step, tread'	*enzak
	badʒaw	ba**r**aw	'to hit'	*bazaw
b.	**d**əpak	**d**agpak	'loud slap'	*dagepak
	lu**d**ah	lu**d**aʔ	'to spit, saliva'	*ludaq
	tu**d**uŋ	tu**r**uŋ	'head overing'	*tuduŋ
	di**d**is	**d**i**r**is	'civet cat'	*didis

In set (a), Malay *dʒ* corresponds to Cebuano *r* intervocalically and *d* elsewhere. In set (b), Cebuano shows the same reflexes corresponding to Malay *d*. Examination of a larger dataset reveals that there is no principled way to account for the Malay distinction between *dʒ* and *d* as a principled phonological alternation. This leads to the view that an ancestral language that gave rise to Cebuano (and nearly all Philippine languages) underwent a merger of these two voiced coronal consonants (PMP *z,*d > *d). The other possibility, that an earlier *d underwent an unconditioned (i.e. arbitrary) split in Malay to become *d* and *dʒ* is discounted by the Neogrammarian hypothesis, which posits that all sound change is regular. The deciding factor in this case is the fact that *z and *d are differentiated in many widely separated languages throughout the Austronesian area. The unlikelihood (or impossibility) of an unconditioned split is thus compounded by having to have occurred in multiple languages independently along precisely the same lines.

We can turn now to the Cebuano distinction between *d* and *r* and ask whether this too may reflect an older phonemic distinction that had been merged in Malay. Contrary to the Malay contrast between *dʒ* and *d*, the distribution of *d* vs. *r* is completely predictable, with *r* only appearing intervocalically and *d* appearing everywhere else. This is therefore simply an allophonic alternation rather than an inherited contrast.

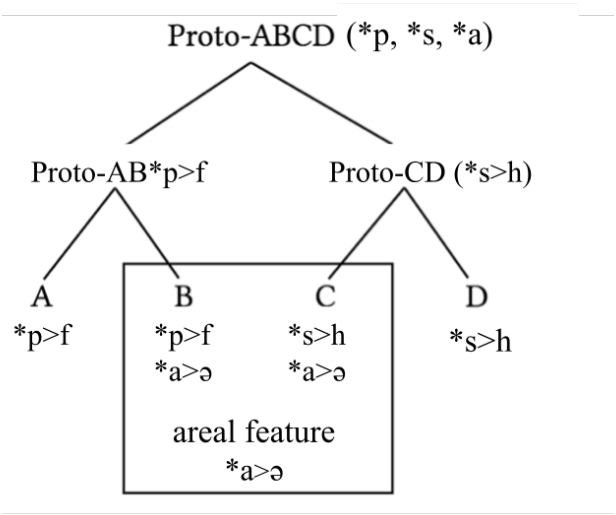
The great advantage of mergers as phylogenetic signals is that they cannot be undone. For instance, once the ancestor of Cebuano and other Philippine languages merged PMP *z and *d, there is no plausible way for a later generation to recoup the

earlier distinction through a subsequent change. Nonetheless, mergers are not immune to the possibility of horizontal spread that most sound changes generally face.

As is often noted in the literature, the phylogenetic trees which result from subgrouping arguments may represent an atypical scenario in which languages split off from each other cleanly and proceed to develop in isolation. We know that the process of linguistic and social divergence is often far messier than this model suggests yet it is a useful fiction which allows us to reconstruct history when such splits do occur. Once the unambiguous cases of branching are recognized, we are also able to confidently identify a residue of changes which have proceeded without such clean breaks (see THE FAMILY TREE MODEL). The crux of the issue is that traditional subgrouping assumes that the transmission of languages and all their features is essentially vertical, from one generation of a language community to the next. However, linguistic change can also be propagated "horizontally" via peer-to-peer transmission across language communities, especially in scenarios with widespread multilingualism or intermarriage (cf. Aikhenvald and Dixon 2001). This is easy to imagine when it comes to the lexicon, as words (by definition) are independent units that are, in principle, free to travel. It may be harder to imagine borrowing a grammatical pattern, an abstract process or a bound morpheme across languages and indeed "grammar", writ large, has often been considered more stable than words for precisely this reason. This is, however, only true under a very narrow definition of language contact, where the loan source is an elite language that only a small portion of the community is fluent in. The French influence on English resulting from the Norman Conquest is often conceived of in this way, with famous lexical doublets consisting of a high register word deriving from French and a lower register word continuing a Germanic etymon (*mutton* vs. *sheep*, *beef* vs. *cow*, *pork* vs. *pig*, etc.). But outside of western Europe, we commonly find extended periods of multilingualism between diverse linguistic groups in far more symmetrical relationships.[i] In such contexts, the canonical contact effect is not borrowing of prestige vocabulary. Rather, the observed effects are more akin to the influence of a speaker's first language (L1) on their second language (L2), which are typically more far-reaching. It is, in fact, difficult to *prevent* syntactic influence from L1 to L2 on the individual level and we thus expect communities with widespread intermarriage or those that have undergone language shift to show at least as much syntactic borrowing as lexical borrowing. While this should not be controversial, we still find a bias in the literature to view contact as primarily a lexical phenomenon and only secondarily as a grammatical and typological one (but see Weinrich 1968, Thomason & Kaufman 1988, Matras 2007, for more holistic views).

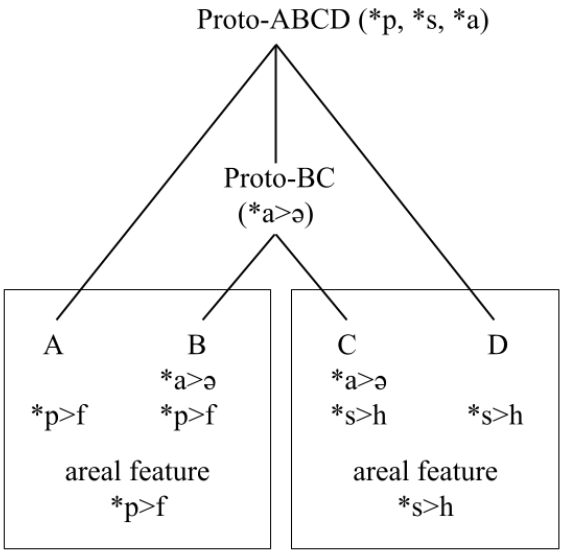
Features that have been transmitted horizontally across language communities can be represented as layers over a traditional phylogenetic tree, as in (3), where the historical change of word-final *a>ə crosscuts languages belonging to two separate branches of a family. The box in (4) represents an areal or "polyphyletic" grouping that does not comport with the hierarchical descent structure represented by the tree.

(4)



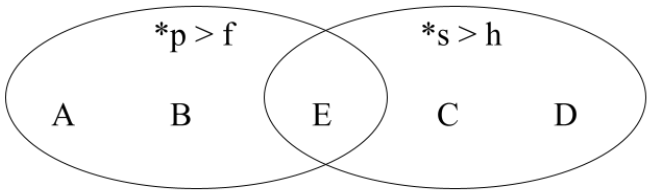
Once it is clear that certain changes support incompatible subgroupings, only detailed knowledge of the languages in question and their neighbors will be able to disentangle inherited innovations from horizontally transmitted ones. For example, the subgrouping in (5), where *a > ə is interpreted as an inherited feature, diagnostic of a subgroup Proto-BC, could be considered less parsimonious than that in (4) by virtue of positing two areal features (*p>f and *s>h) rather than one (cf. Wichmann 2010:73). Nonetheless, it must be taken into account when evaluating the full range of possibilities.

(5)



A similar dilemma is encountered with languages that display a set of innovations that define nearly exclusive subgroups, in addition to a small number of languages where these innovations overlap. These scenarios are the raison d'etre of the wave model (Schmidt 1872, see WAVE MODEL, this volume), which conceives of change as propagating primarily across communities rather than by descent, and which would represent the situation as in (6).

(6)

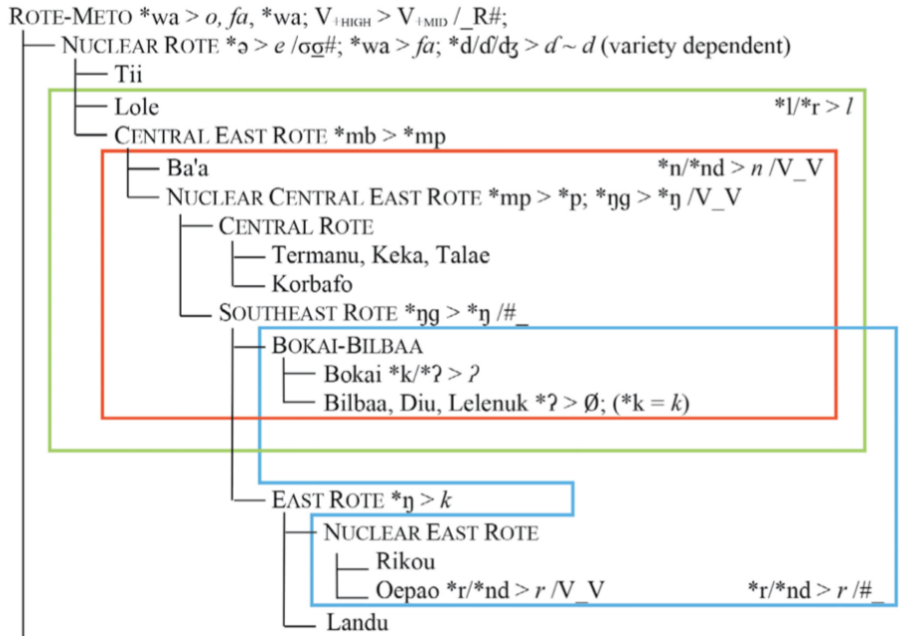


It is a misconception that transmission of linguistic features that crosscut standard phylogenetic groups, as the areal features above, can only come about through highly distinct and unusual means. Ross (1988) argues that areal features may result from the same natural processes that give rise to inherited changes when we consider the language community as a complex social network with partially overlapping areas of innovation. When such a dialect network splits and continues to differentiate, it is very possible that features become distributed in ways that are impossible to capture in a simple family tree. Ross (1988 et seq) terms such a language relation a linkage, which can be represented as a series of overlapping innovations, much as in (6).

François (2014) argues for treating linkages as the norm rather than an exceptional overlay on a traditional family tree (cf. Dixon 1997). He proposes that the family tree is just one special case of a linkage or wave-like distribution, specifically, a wave model in

which all innovations happen to be nested. This challenges the classical definition of a genealogical subgroup, which François defines as “a group of languages whose ancestors participated together in the diffusion of one or several linguistic innovations, at a time when they were mutually intelligible”. This not only implies that subgroups can intersect, it essentially eliminates the differences between inheritance and borrowing. It is true that the natural limits of language contact and horizontal transmission are understood today to be far greater than previously imagined and that the historical effects of multilingualism are still not fully appreciated. Nonetheless, the reality of borrowability clines is well established and different types of transmission and change most likely indicate different relations. While strict stammbaumian subgrouping erases areal effects, strict wave-based subgrouping fails to differentiate easily borrowable innovations (e.g. lexemes) from those which are less likely to cross language boundaries (e.g. inflectional morphology, complex, opaque, and unexpected sound changes). Careful analysis by semantic field also shows that different domains appear to reflect different relations as well. Thus, while the renewed interest in wave theory may serve as a corrective to arboreal dogmatism, it seems that a marriage of the two approaches is justified. In practice, we often find a happy medium in the form of a stammbaum overlaid by waves, as exemplified by Edwards’ (2021) phylogeny of the Rote-Meto languages of the Timor area, a portion of which is shown in Figure 1. Cross-cutting innovations are represented here as boxes over a tree diagram and subgroup defining innovations are shown next to the name of each reconstructed protolanguage. This type of visualization is ideal as it shows in a single diagram the strength of the evidence for each subgroup in comparison with the distribution of each areally defined change. As noted earlier, this is important in evaluating which changes are best analyzed as subgroup defining and which are areal, as diagrams such as those in Figure 1 are merely hypotheses in competition with many other possibilities. Further discussion of the integration of areal features in subgrouping can be found in THE FAMILY TREE MODEL this volume, Jacques and List (2019), and references therein.

Figure 1.A portion of the Rote-Meto tree overlaid with areal changes (Edwards 2021)



2.1 Domains of evidence

Although sound change is by far the most studied and commonly employed type of subgrouping evidence, it is known that certain types of sound changes are susceptible to horizontal spread and that the strongest subgrouping arguments are based on innovations across multiple linguistic domains.

Morphology has long played a prominent role in subgrouping (cf. Bopp 1816) although not all types of morphology carry the same probative value. Clearly, morpheme types that are easier to borrow only provide weak evidence for subgrouping. In this respect, bound morphology is generally understood to be harder to borrow than free morphemes or whole words, although it is clear that bound morphemes are routinely borrowed as well.[ii] Within the realm of bound morphology, inflectional morphemes that form part of a paradigm are less likely to be borrowed than derivational morphemes. Morphemes which convey concrete and discrete meanings are also more easily borrowed than those with abstract or purely grammatical functions. Several clines of borrowability have been posited in the literature, e.g. Matras (2007), as shown in (7).

(7)a. Free lexemes > bound derivational morph > bound inflectional morph
independent morphology > paradigmatic morphology
semantically discrete & concrete > semantically abstract & functional

b. affixal morphology > non-affixal morphology

c. regular morphology > irregular morphology

Accordingly, those elements further to the right of these scales must be seen as more reliable indicators of phylogenetic relations than those to the left. In the Austronesianist literature, Zorc (1974b:426, 1977, 1986) establishes that function words are far better predictors of genetic relations than basic lexical vocabulary and that even within basic vocabulary, certain lexical items are extremely stable throughout Austronesian (e.g. 'tongue', 'hand', 'heavy', 'eat') while others are much less so (e.g. 'all', 'fat', 'lightweight') (cf. Tadmor et al. 2010, Haspelmath et al. 2009 and Seifart 2020 for a global overview).

Free pronouns have generally been considered as part of functional/paradigmatic morphology and have thus been frequently employed in subgrouping arguments. Thomason and Everett (2005) show, however, that many clear cases of pronoun borrowing exist throughout the world, with certain geographic regions (Southeast Asia prominent among them) being exceptionally rich in examples. Where pronouns have proved especially useful is in the recognition of old families where little other evidence exists (e.g. Ross 2001a, 2005), as we do not expect pronoun borrowing to affect a wide range of geographically dispersed languages in the same way. Nevertheless, innovations in pronominal systems can provide strong subgrouping arguments so long as the possibility of borrowing can be minimized.[iii]

Syntactic evidence, in principle, may be recruited for subgrouping purposes in just the same way as sound change and morphological change. However, in practice, syntactic evidence is employed far less for several reasons. First, the study of syntactic structure simply lagged far behind the study of sound structure and word structure in the development of historical linguistics. Classical descriptive grammars often contained little to no information about syntax and were chiefly concerned with phonology and morphology, often to the extent that "grammar" was taken to be synonymous with word structure. Secondly, it is only with a well developed theory of syntax and, most importantly, syntactic change that grammatical patterns can be argued to be innovatory. To take a simplistic example, given a set of languages some of which show basic SVO word order with others showing VSO word order, it is not immediately clear which order should

be considered innovatory as SVO>VSO and VSO>SVO both appear to be well-attested changes. While the directionality of change can often be deduced from historical records (Cumming 1991) or even synchronic syntax (Aldridge 2010), attempts at positing sweepingly general unidirectional changes in syntax as in Givon (1979) have generally met with failure. Campbell & Harris (1995:332) give the example of Klimov's hypothesis that alignment systems only develop in the direction ACTIVE > ERGATIVE > NOMINATIVE, and show that there is much good evidence to suggest that ergative and active alignments may develop from nominative ones. As Campbell & Harris (1995:343) also note, there are few good candidates for context-free unidirectional changes in syntax, but this is also true for sound change, where the unidirectionality of debuccalization (*s>h, *k^h>h, etc.) represents the exception rather than the norm.

Harrison (2003) expresses a wholesale critical view towards syntax as evidence for a phylogenetic relation but this appears to be based on an idiosyncratic notion of syntax as merely an abstract pattern without relation to particular morphemes showing regular phonological correspondences. To be sure, simple word order patterns are of this nature but most syntactic arguments for subgrouping are *morpho*syntactic, implicating particular morphemes and their historical reinterpretation rather than pure ordering relations. The weakness of ordering relations as subgrouping evidence has been well established and the Austronesian family offers many examples of how changes in word order (e.g. VSO > SVO, SVO > SOV, Adj N > N Adj, Dem N > N Dem, N Poss > Poss N) follow an areal pattern rather than a genetic one (Donohue 2007, Kaufman 2009b). On the other hand, the reinterpretation of a particular subordinate clause type as a matrix clause predicate (as in Ross 2009 and Aldridge 2016) or the reinterpretation of a stative intransitive affix as a transitive one (as in Chen et al. 2022) are robustly syntactic phenomena, but at the same time rooted in the behavior of a limited number of morphologically signaled derivations. The details of such developments can be sufficiently "surprising" (to use Harrison's term) as to make for a strong subgrouping argument. Furthermore, if such a pattern spread areally then we may expect it to be accompanied by lexical loans, whose spread typically precedes that of grammatical morphology and syntactic patterns (although see Ross 2001b on "metatypy", where syntactic change proceeds without extensive lexical borrowing).

Grammaticalization phenomena (Hopper & Traugott 2003, Heine & Kuteva 2002) often provides strong evidence for directionality in morphosyntactic change. The core thesis of grammaticalization is that functional morphology originates from free lexical items via a process of phonological reduction and semantic bleaching. Thus, if we find applicative or case marking affixes in certain languages whose cognates in other languages correspond to independent adpositions (i.e. free lexemes), we can be confident in treating the affixes as innovatory due to universal diachronic tendencies. The unidirectionality of such changes has been criticized by Janda and others, and many exceptional "degrammaticalizations" have been documented (cf. Norde 2009), yet the overall force of the argument is generally seen to hold. Not all syntactically based arguments need to rely on grammaticalization, as directionality can be inferred by other means. What does seem necessary, however, is a syntactic framework that allows for precise cross-linguistic generalizations and predictions.

An enduring difficulty in morphosyntactic reconstruction is discerning the meaning of zero, that is, whether the lack of a particular linguistic feature (morphological paradigm, syntactic pattern, etc.) in one language represents a continuation of an earlier stage before the feature had developed elsewhere or an innovatory simplification. This problem has arisen prominently in at least two cases in Austronesian. The lack of one voice paradigm in the Formosan languages Rukai and Tsou, and its restriction to relative clauses in a third Formosan language, Puyuma, is taken to reflect the primordial state of affairs by Starosta et al (1982), Ross (2009) and Aldridge and Yanagida (2021) but interpreted as a secondary loss by Chen (2017) and Blust & Chen (2017). Similarly, the lack of a full person marking system in certain Celebic languages is taken by Mead (2002) to reflect earlier stages of its accretion but interpreted by van den Berg (1996) to reflect

secondary loss. How can such cases be adjudicated? Blust & Chen (2017) emphasize that absence of evidence is not evidence of absence; if a language lacks a particular pattern, it could have possessed it at an earlier stage and subsequently lost it. Recall, though, from (2), that without Occam's Razor many types of otherwise plausible innovations could be posited thus rendering our subgrouping moot. On this basis, parsimony is a necessary component of the comparative method and privileges a scenario in which the lack of a pattern represents a retention rather than the result of development and subsequent loss, all else being equal. But all else is rarely equal and there are diverse factors and sources of evidence to consider in the domain of morphosyntax. The burden of proof should lie with the less parsimonious scenario in such cases, as we may expect clues to an earlier pattern in frozen morphology. In the case of Celebic person marking, we see how cross-linguistic evidence can be brought to bear as well. Several subgroups of Sulawesi and Sumatra show a partial set of pronominal verbal prefixes that typically express a transitive agent. In Malayic languages, where we have the great advantage of epigraphic evidence, we can be relatively certain that the development to a full paradigm took place stepwise through the accretion of first and second person markers followed by third person markers. Old Malay shows no prefixes at all; classical Malay shows prefixation of first and second person agents, while several related modern languages show prefixing for all persons. Meanwhile, on Sulawesi, we observe that incomplete prefixing paradigms either contain just first person forms, or first person and second person forms. In no case do we find a language with third person prefixes that lacks first and second person prefixes. Lacking any direct evidence suggesting that third person prefixes are lost before first and second persons, it would seem that the partial agreement patterns of Celebic languages represent stepwise accretions on the way to a full paradigm rather than loss from a full paradigm (Wolff 1996, Kaufman 2014).

It is difficult to make blanket statements regarding the relative strength of innovations in morphology versus syntax versus the lexicon. While there is general agreement that functional morphology is harder to borrow, each case tells its own story. For example, Beck (2023), examining the Totonacan languages of Mexico, argues that changes in the inflectional system which at first sight appear to be strongly diagnostic for subgrouping purposes, appear on closer inspection to have spread horizontally due to details in their distribution and use. Here, he claims, it is the lexical innovations that are the better indicator of phylogenetic relations, contrary to the general case. Needless to say, the most convincing subgrouping arguments are based on a range of evidence from different areas of language. As Brugmann (1884, translated by Dyen 1953: 580) himself states: "it is not a single or a few linguistic phenomena appearing in two or several areas at the same which furnish a proof of closer community but only a large mass of agreements in sound, flectional, syntactic and lexical innovations, the large mass of which excludes the thought of accident."

3. Case studies

It is only by comparing secure subgroups with controversial ones that we can gain an appreciation of the quantity and quality of evidence that is generally deemed convincing. I review several non-contiguous subgroups of the MP branch of Austronesian in which the quality of the innovations are seen to outweigh geographical considerations. These cases demonstrate most clearly the triumph of the comparative method in disentangling linguistic history, making an important contrast with similarity-based methods, discussed in §4.

All the following cases are drawn from the MP languages, which are shown in Figure 2 (adpated from Ross 2008), a family tree argued for by Robert Blust in a series of publications (see summary in Blust 2013). Among the higher level subgroups shown, it is only Oceanic and MP that are entirely undisputed. As indicated by the question marks and

the “linkage” appended to Central MP, a number of subgroups defended by Blust have been argued to instead represent linkages arising from extensive and prolonged contact.

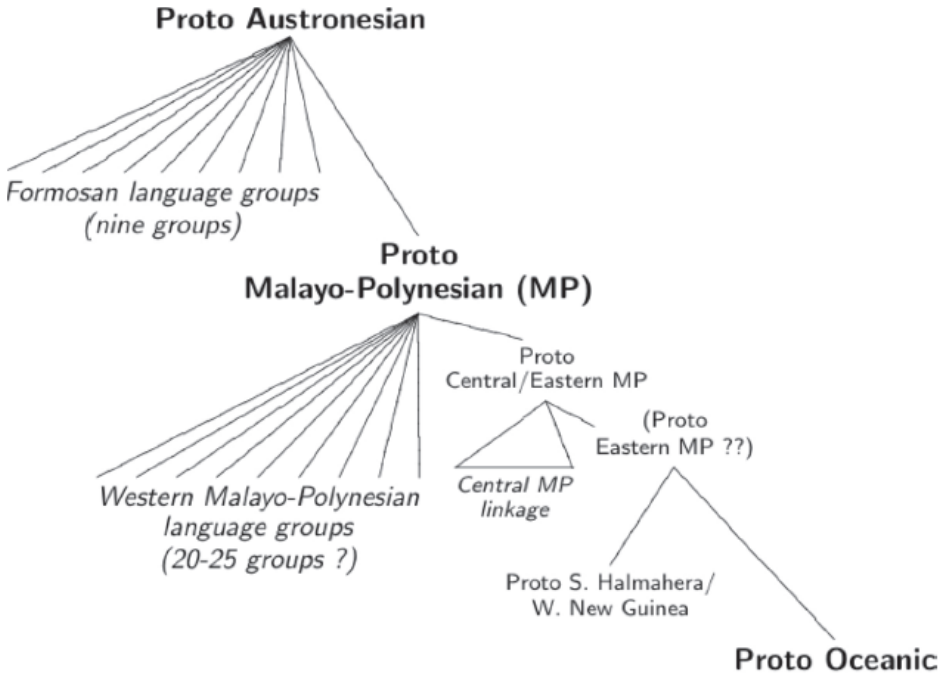


Figure 2. The major branches of the Austronesian family (adapted from Ross 2008)

The Malayo-Polynesian (MP) languages constitute all those Austronesian languages spoken outside of the island of Taiwan (with the exception of Yami, a member of the northernmost Batanic subgroup of MP located on Taiwan’s southernmost island). Because of the size and enormous geographical breadth of the Austronesian family, it is difficult to imagine that all languages outside of Taiwan (comprising the vast majority of Austronesian languages) would all share a common set of innovations not found in those Austronesian languages within Taiwan, commonly known as the Formosan languages. Until Dahl (1973) and Blust (1977), a working assumption of earlier scholars, based purely on a geographical bias, was that the Formosan languages, being neatly contained on a single island, formed a subgroup amongst themselves. Blust (1995, 1999:55-57) shows that phonological arguments adduced by earlier scholars in favor of such a Formosan subgroup all drew upon conservative aspects of individual phonologies, e.g. sibilant reflexes of *S, uvular reflexes of *q, and maintenance of contrasts that have been lost outside of Taiwan. Once such features are recognized as retentions, support for an exclusive Formosan subgroup evaporates. It is now accepted universally (or very nearly so) that the Formosan languages form several primary branches of the Austronesian family tree.[iv] The most prominent changes that support the large MP subgroup, comprising all the Austronesian languages outside of Taiwan, are shown in (8)-(12).

(8) Irregular syncope and C-deletion
PAN *paŋudaN ‘pandan’ > PMP paŋ(e)dan
PAN *biRbiR ‘lip’ > PMP *bibiR

(9) Irregular *S > ∅ in several etyma
PAN *Sepat > PMP *epat

(10) Regular mergers

*C, *t > *t (where *C was likely [ts] or [tʃ])

*S, *h > *h (where *S was likely [s]).

*N, *n > *n (where *N was likely [n̪], [lʲ] or [ɬ])

(11) "Politeness shift" (Blust 1977)
replacement of *-Su 2SG.GEN by *-mu 2PL.GEN

(12) *S metathesis

PAn *bukeS	> PMP *buhek	'hair'
PAn *CaqiS	> PMP *tahiq	'sew'
PAn *tapeS	> PMP *tahep	'winnow'

The metathesis of *S with a preceding stop, shown in (13), followed by (or simultaneous with) PAn *S > PMP *h is one of the strongest pieces of evidence for PMP, despite applying only sporadically to several words. Firstly, only reflexes of the metathesized forms *buhek, *tahiq and *tahep are found outside of Taiwan and only reflexes of the unmetathesized forms, *bukeS, *CaqiS and *tapeS, are found in Taiwan. Secondly, metathesis of non-adjacent segments that are not typically prone to metathesis (cf. Blevins and Garrett 2004) is extremely unlikely to have occurred several times independently.[v] Thirdly and most importantly for subgrouping purposes, we can be confident that the direction of the metathesis was *CVS > *hVC and not the other way around because, alongside cases of metathesis, we have many instances of medial *h in PMP that correspond to medial *S in PAn. Examples like the near minimal pair of 'head hair' and 'drunk' in (13) show that when metathesis applied it shifted *S to a preceding position in the word but left word-medial *S as is. Finally, the contemporaneous change PAn *S > PMP *h is one whose directionality is all but certain, as fortition of /h/ to a sibilant is nearly unknown as a regular sound change.

(13)

PAN *bukeS	> PMP *buhek	'head hair'
PAN *buSuk	> PMP *buhuk	'drunk'

The strength of the preceding changes, in addition to a large reconstructed vocabulary, has allowed us to deduce sheerly on the basis of linguistic evidence that Taiwan was the original home of the Proto-Austronesians, with a single breakaway group ultimately giving rise to all those Austronesian languages outside of Taiwan. However, the combined strength of (8)-(12) is exceptional among high level subgrouping arguments in Austronesian. But even weaker changes, which may be poor in probative value on their own, can gain strength in numbers.[vi]

The Taiwanese homeland hypothesis has since found strong support through extralinguistic evidence from genetics and archeology (see Bellwood 2017 for a summary).

3.1.2 Oceanic

The earliest hypothesis of a subgroup containing over 450 languages of the Pacific is due to Dempwolff (1937), as one result of his overarching reconstruction of Proto-Austronesian. Later research by a number of scholars (see Lynch et al 2002) further substantiated the phylogenetic status of the Oceanic languages as a lower level daughter of Proto-Malayo-Polynesian, as seen above in Fig. 1. It is now agreed upon that most of the Pacific islands were first inhabited as part of the Austronesian expansion from the area of eastern Indonesia/western Papua (Bellwood 2017) although many of the Austronesian languages of this region have been in long standing contact with non-Austronesian

languages, often referred to as “Papuan” languages (although this is a purely areal term and not a phylogenetic one).

Despite long and complex histories of contact and movement in this region, evidence for a discrete Oceanic subgroup is surprisingly strong. The clearest innovation is a large set of phonological mergers, including *b/p > *p, *mp/mb > *b, *g/k > *k, *ŋk/ŋg > *g, *d/r > *r, *e/aw > *o. Some of these, such as *b/p > *p, are unknown elsewhere in the Austronesian family and are thus of great diagnostic value. The fact that these sound changes contain multiple unrelated processes (devoicing, postnasal voicing, cluster simplification, rhoticization of *d and monophthongizaton) further strengthens their value. Grammatically, there is great diversity among Oceanic languages although much of this appears to have been due to contact with non-Austronesian languages after the break-up of the family. Ross (1988:119) notes that morphological simplification (e.g. the replacement of aspect inflection by independent auxiliaries) must have taken place independently across a wide area of the western Oceanic area as well as in other branches of Central-Eastern Malayo-Polynesian (CEMP) languages. There exist innovations in the development of a verbal person marking system and noun classification system that can be reconstructed to Proto-Oceanic and may serve as independent subgrouping evidence (Lynch et al. 2002:68), although the uniqueness of these developments to the Oceanic subgroup is unclear.[vii]

The existence of an Oceanic subgroup has not faced serious challenge, due in large part to the exceptionless set of mergers and extensive set of unique lexemes (Ross, Pawley and Osmond 1998 et seq).

Ross’s (1988) internal subgrouping of Oceanic remains a model of rigor and breadth, in taking evidence from comparative phonology, morphology as well as syntax to establish six higher level groupings. Moreover, Ross successfully integrates linkages and other apparent wave-like distributions into a moderately conservative stammbaum-based phylogeny. The relatively clean break between Oceanic and non-Oceanic languages with regard to the subgroup’s defining innovations eliminates the possibility that the Oceanic subgroup is merely a sprachbund. However, within Oceanic itself, sprachbund effects, linkages and other “network effects” abound, posing many difficult questions for lower level subgrouping.

3.1.3 Malagasy as a member of the Southeast Barito subgroup

Malagasy, the native language of Madagascar, was understood quite early to be belong to the Austronesian family but its relation within the family was largely conjecture until the mid-20th century (Adelaar 1995). While both Humboldt (1836-9) and van der Tuuk (1865) had posited close relationships to Philippine languages and Batak, respectively, their hypotheses were not based on exclusively shared innovations but rather morphological and syntactic similarities that are now understood to be retentions from PMP. Their arguments for a special connection between Malagasy and Philippine languages are thus now considered invalid. It was Dahl (1951) who first proposed that Malagasy subgroups together with the far flung Maanyan language of Borneo on the basis of innovations, a thesis that has since been refined by Mahdi (1988), Adelaar (1989, 1995, 2010, 2012), among others.

Although some of Dahl’s (1951) arguments have now been weakened or invalidated, either as retentions, early borrowings from Malay, or ancient sprachbund effects (cf. Adelaar 2021, Smith 2018), the remaining evidence is sufficiently strong as to yield a consensus opinion that Malagasy belongs to a far-flung subgroup all of whose other members are found in south Borneo. This result, which could not have been anticipated on the basis of geography, material culture, or the ethnographic record, was bolstered by Hudson (1967), who examined the Barito languages in new detail and offered arguments for several low level Barito subgroups, which in turn led to the inclusion of Malagasy as part of his Southeast Barito subgroup. The sound changes supporting this classification include the shared set of innovative lenitions: PMP *d > PSEB *r, PMP *b > PSEB *w, PMP *s

> PSEB *h, PMP *R > PSEB *y (with PSEB *h and *y further developing into Malagasy ∅ and z, respectively), in addition to morphological and lexical evidence first adduced by Dahl (1951).

Ignoring the overarching morphosyntactic typology and focusing on sound change and specific morphological innovations revealed that Malagasy, a language which shows many of the same typological features as Philippine languages (e.g. a rich inventory of voice morphology that promotes an argument to subject, verb-initial word order, case marking, second-position clisis) is more closely related to Maanyan and other languages of the Southeast Barito group, which are all SVO languages with far simpler morphological profiles and no case marking on full noun phrases. On the other side, Malagasy has undergone a restructuring of its phonology due to influence of Bantu languages. In this regard, Adelaar (2012) counts final open syllables, lenition of stops and cluster reduction and, less directly, the stress pattern, as having a possible origin in Bantu contact. While Malagasy morphosyntax is strikingly conservative, certain of its features also appear to betray a Bantu influence, such as the three-way tense system and, more transparently, the use of a Bantu class prefix in a diminutive function.

Despite the typological discrepancies in the phonology, the subgrouping of Malagasy with the Southeast Barito languages is now established. Not only does this subgrouping shed a vast light on the origins of the Malagasy, it also holds a fascinating implication for the history of Borneo and its languages. Namely, it makes clear that wide-ranging processes of morphological simplification swept over Borneo after the ancestors of the Malagasy had left, thus obscuring their typological tracks (Dahl 1951, Adelaar 1995, 2010, Blust 2013:70). The question of what could have triggered these changes over such a wide-ranging area remains open. Malagasy thus represents a significant triumph of the comparative method in elucidating subgrouping and population movement within the Austronesian world. It led to a conclusion that, firstly, has no other plausible explanation and, secondly, could have only been established with confidence using the comparative method, especially given the great distance between Malagasy and its closest congeners.[viii]

3.1.4 Chamic as a member of Malayo-Chamic subgroup

Just as Malagasy had undergone contact with Bantu languages leading to an open syllable typology, we find the Chamic languages, geographical outliers of the Austronesian family located in mainland Southeast Asia, underwent radical restructuring via contact with Austroasiatic languages. The Chamic languages likely descend from the language of Champa, a kingdom that dominated maritime trade along the Vietnamese coast from 2nd to 14th century CE. As noted by Thurgood (1999:31), Chamic was thought originally to be an Austroasiatic language by Schmidt (1906) and Sebeok (1942), that is, more closely related to Vietnamese than to any Austronesian language. It is now undisputed that Proto-Chamic and Proto-Malay were very similar languages, sharing a set of unique innovations that justify the phylogenetic group Malayo-Chamic, including PMP *h > ∅, *q > *h, PMP *w > ∅ / #_, the mergers PMP *j, d > d and PMP *R, *r > *r, as well as lexical and morphological evidence.

Thurgood (1999) shows how the Chamic languages converged typologically with their Mon-Khmer neighbors over centuries of close contact, beginning with an iambic stress pattern that led to the development of a sesquisyllabic typology that included consonant clusters typical of Mon-Khmer languages. This development towards Mon-Khmer phonological typology culminates in the development of monosyllabic roots and tonogenesis, as can be seen in the comparisons between Malay and Phan Rang Cham in (14).

(14)

Malay	Phan Rang Cham	
pohon	pʰun	‘tree’
baharu	frew	‘new’
bəri	prày	‘give’
bəras	pràh	‘rice’
mari	maay	‘come here’
toloŋ	trun	‘help’

Morphologically, Chamic languages have also moved in a strong analytic direction, corresponding with their Mon-Khmer neighbors, from whom they adopt SVO word order as well as many other syntactic features. Many Chamic languages have also undergone massive lexical replacement, with up to half of their vocabulary borrowed from surrounding Mon-Khmer languages.[ix] The examples in (15) and (16) hint at the large scale morphosyntactic changes that have pulled the Chamic languages away from Malayic typology.

(15)a. *Phan Rang Cham* (Blood 1978:43)

tə́hlaʔ naaw thaŋ aay hu laayʔ
I go house elder.brother able Q
‘Can I go to your house?’

b. *Malay*

Boleh saya pergi ke rumah kakak?
can 1sg go to house older.both
‘Can I go to your house?’

(16)a. *Phan Rang Cham* (Thurgood 2005:509)

min ətày oh diiʔ cìŋ pɛʔ ka əmɛɛʔ baŋ
but y.sibling NEG climb able pick for mom eat
‘but younger sibling can’t climb up to pick it for Mom to chew.’

b. *Malay*

tapi adik tidak bisa naik untuk məm-[p]etik-nya buat di-makan ibu
but y.sibling NEG able climb for AV-pick-3s.GEN for PV-eat mom
‘but younger sibling can’t climb up to pick it for Mom to chew.’

Note that there are relatively few shared cognates between the two languages in the examples and that Phan Rang Cham follows Vietnamese word order in postposing modals such as *hu* and *cìŋ* ‘able’, as well as showing serial verb constructions. Furthermore, the actor voice/undergoer voice distinction which remains an important part of many Malay varieties has been completely eliminated in Chamic, together with most verbal morphology (Thurgood 2005:507). However, the subgrouping of Chamic and Malay via the comparative method, on the basis of the aforementioned sound changes and other innovations, is strongly vindicated by a Chamic inscription from the 4th century CE, the earliest inscription of any Austronesian language. The inscriptional evidence provides the missing link between a more Malay-like language and its modern descendants, which appear typologically similar to Mon-Khmer languages.

A relatively small subgroup containing Chamic and Malayic languages is now well accepted and was only deduced by a careful analysis of shared innovations. The great challenge of subgrouping Chamic lay less in discovering hidden data tying it to Malay, as that data was known from the earliest studies, but rather in assiduously ignoring the

overall phonological, morpholological and syntactic typology, as well as half the lexicon, all of which was due to contact with local non-Austronesian languages (cf. Donohue et al. 2008, Donohue and Grimes 2008, Donohue et al. 2011 for similar arguments on the eastern side of the Austronesian family).

3.1.5 Abaknon as a Sama-Bajaw language

Abaknon (also known as Inabaknon and Capuleño) is the language of Capul island, lying between the large Philippine island of Samar and the southern tip of Luzon. Capul island is surrounded by Central Philippine languages, specifically ensconced within a Bisayan language area. Today, the Abaknon people are difficult to distinguish from their Bisayan neighbors culturally, yet their language only bears a superficial resemblance to the surrounding Waray language. Abaknon has been shown to belong to the Sama-Bajaw subgroup, a group of languages that are spoken mostly by nomadic "sea gypsies" inhabiting a wide area spanning from eastern Indonesia to the Sulu archipelago and Borneo. Capul island represents the northernmost attested branch of the subgroup (Pallesen 1985, Kaufman 2024) and is separated from the closest Sama-Bajaw language by over 500 kilometers. Abaknon sits in the middle of the Greater Central Philippine (GCP) zone (Blust 1991), whose defining features include PMP *R > g and lexical replacements such as PMP *wahiR > PGCP *túbig 'water' and PMP *Rumaq 'house' > PGCP *balay 'house'. Abaknon, uniquely for this region, maintains reflexes of both PMP *Rumaq (*ruma*ʔ) as 'house' and *wahiR (*buwahi*ʔ), the first of which also reflects *R > r, unlike surrounding Philippine languages. While Abaknon lacks the lexical innovations that define the surrounding GCP languages, it contains lexical innovations such as *aʔa* 'person' and *-bi* 2PL, that are unique to Sama Bajaw languages, in addition to consonant gemination following historical schwa as in *təlu > tallo 'three', and other sound changes that are unknown in the surrounding area.

Abaknon morphosyntax has not been sufficiently studied but shows influence from Waray in the functional morphology while maintaining what appear to be Sama-Bajaw innovations. The mixed history of the language can be seen in (17)-(18), where underlining indicates inherited morphemes and bold typeface indicates borrowings from various sources (Waray, Spanish and English). Plain typeface is used for words whose origin is ambiguous. These typical examples give a rough idea of how much the language has been affected by contact with a wide range of languages, both Austronesian and Indo-European (i.e. Spanish and English).

Abaknon (Jacobson 1999)

(17) Bawa-ko iya pan **huspital basi' manggad** pa **kon** an-halap iya
carry-1S.GEN 3S.NOM to hospital for have.chance still if AV-good 3S.NOM
'I will take him to the hospital in order that he might get well'

(18) Ma-tappo' i sanga-na si **rimas**
ABL-break NOM branch-3S.GEN OBL breadfruit
'The branches of the breadfruit are breakable.'

Pallesen (1985:36) calculates that Abaknon shares 41% of its vocabulary with its Bisayan neighbor Waray and has replaced approximately 17% of its basic vocabulary while Blust (2007:79) calculates that at least two thirds of the vocabulary is borrowed. In some cases, we find doublets in the basic vocabulary with one member displaying expected sound changes and the other an apparent Waray borrowing, as in PMP *baqeRu > *baha*ʔ*o* 'new' (with Sama-Bajaw *R > h) next to *bag*ʔ*o* (with Bisayan *R > g). If the vocabulary of Abaknon shows a near even split between a Sama-Bajaw origin and a Bisayan origin, why not

consider Abaknon as a member of the Bisayan subgroup influenced by a Sama-Bajaw language? Here we rely on our understanding of what linguistic elements are more susceptible or more impervious to horizontal transmission. As discussed earlier, established clines of borrowability exist on several dimensions, including lexical categories. While there is some disagreement for minor categories (cf. Haugen 1950, Muysken 1981, Matras 2007), there is consensus over most of the major lexical categories and types shown in (19), where each scale goes left to right from most borrowable to least borrowable (Matras 2007:61).

(19) **Category**: Nouns, conjunctions > Adjectives > Verbs > Prepositions > Pronouns
Morphological Type: derivational morphology > inflectional morphology
Boundedness: free forms > bound forms

Even within the minuscule language sample given above in (17)-(18), we find that it is precisely those categories that are most easily borrowable that have a non-Sama-Bajaw etymology, as seen in (20), where the ratios are shown to the right.

(20) **Nouns:** *ismaglir, baligya*ʔ*, sundalo, huspital, rimas* (0/5)
Conjunctions/subordinators: *kon, basi'* (0/2)
Verbs: *bawa, anhalap, matappo* (3/3)
Prepositions: *pan, si* (1/1)
Pronouns: *iya, -na* (2/2)

As this pattern appears to hold over the lexicon as a whole, the likelihood of Abaknon being a Bisayan language that was later subject to Sama-Bajaw influence is highly implausible. Such borrowability clines are crucial to classification and subgrouping when dealing with languages that have long histories of heavy contact, which require teasing apart borrowings from inherited elements on a larger scale than usual.

3.2 Debated subgroups

Standing in contrast to the widely accepted subgrouping proposals reviewed above, there are several MP subgroups whose sole basis is lexical. These subgroups, which include Western Indonesian (Blust 2010, Smith 2017), Greater Northern Borneo (Blust 2010) and the Philippine subgroup (Charles 1974, Paz 1981, Zorc 1986, Blust 2019), have not fared so well under additional scrutiny. Smith (2023) offers a critical look at the first two cases. Here, we focus on the last.

3.2.1 Proto-Philippine

Blust (2005, 2019, 2022) proposes a Philippine subgroup on the basis of a single, common sound change and a large set of lexical innovations. This proposal repays careful study as it best exemplifies the dangers of subgrouping by lexical evidence.[x] The northern border of the putative Philippine subgroup had already been established, as it is the same border between the Malayo-Polynesian languages and the Formosan languages of Taiwan, the latter of which represent several primary subgroups of Austronesian. The challenge of establishing a Philippine subgroup is in its southern border. Earlier attempts at reconstructing Proto-Philippines (PPh) were based on a sample of convenience, often exclusively consisting of languages within the national borders of the Philippines (cf. Paz 1981, Reid 2017, Blust 2019). Based on a seeming lack of purely Philippine innovations, Reid (1981) argued that the Philippine subgroup lacked credible evidence, a stance that

accords well with an "express train" scenario of the Austronesian expansion involving a rapid southwards demographic spread from the initial departure from Taiwan. However, beginning with Zorc (1986) and culminating in Blust (2019, 2022), an impressive number of lexical innovations have been arrayed in favor of a Proto-Philippines, to wit, a total of 1,606 lexemes. Blust (2019) stresses that even if half of this list is whittled down by the discovery of external cognates and evidence of inter-Philippine borrowing, the remaining 800 innovations should be more than enough to substantiate the subgroup.[xi]

In the ensuing debate between Blust (2019) and critics of the Proto-Philippine hypothesis (Liao 2020, Reid 2020, Ross 2020, Chen et al. 2024), we find a large rift in understanding the concept of "negative evidence", which can be compared to the aforementioned debate on the significance of morphological zero in subgrouping arguments. Liao (2020), after pointing out that the single phonological merger that defines the subgroup (PMP *z, *d > PPh *d) does not actually take place in all Philippine languages (in addition to taking place in many non-Philippine languages), claims that the remainder of the proposal hangs entirely on "negative evidence", specifically, the apparent lack of cognates outside of the Philippines for the list of lexical innovations. Blust (2020:472) counters this critique, stating that all proposals are inherently probabilistic and that, while each innovation is weak evidence on its own, it would require falsifying over 1,200 (now 1,600) lexical innovations to refute the hypothesis. This, however, is based on a subtle premise that each of the claimed innovations represents an independent historical fact. In actuality, the distribution of these 1,600 lexemes could be due not to 1,600 individual innovations but to a far smaller number of secondary developments. The most obvious possibility, raised by Ross (2020) and Smith (2017), is that these words spread through early networks within the Philippine zone, despite Blust's precautions to avoid reconstructing words with loan distributions.[xii] Sound correspondences are generally used as a diagnostic to differentiate inherited etyma from loans from related languages. Apparent shared lexical innovations which display diagnostic sound changes are thus often taken as subgrouping evidence. But here it is crucial to separate one-off sound changes, which have a higher diagnostic value, from synchronic patterns that affect loan phonology, which have little diagnostic value. For instance, if a language lacks voiced stops in coda position altogether, coda devoicing cannot be taken as strong evidence for a native etymology, as it is most likely also a part of loan phonology. However, patterns with more complex conditioning might not be incorporated into the synchronic phonotactics of the language and it is these changes which have real probative value. Another inherent difficulty in lexical evidence is its oftentimes uneven distribution throughout the witness languages and subgroups. Liao (2020) notes that none of the original 1,286 proposed PPh innovations are found in all the constituent subgroups while Blust responds that this is simply a natural outcome of the vagaries of "differential retention" (cf. "incomplete lineage sorting" in Jacques and List 2019). The picture is complicated by the fact that, as even Blust (2020:454) admits, there must have been a number of linkages overlaying any putative Philippine subgroup. While we do not necessarily expect the majority of etyma to appear in all subgroups, the typically spotty distributions do raise questions. Blust counters that if we find an etymon with all the expected correspondences in the extremities of the zone and nowhere else, attributing it to a chance resemblance is still highly unlikely; either the word must be inherited or a loan. If there is no attested relationship between two distant languages and nothing in the phonology that indicates a loan, then it could have only been an inherited from a common ancestor. In this way, a lexical reconstruction may find strong support from only two distantly separated witnesses.[xiii] As support, Blust cites Mallory and Adams (2006), who reconstruct a word meaning 'hunger' to PIE despite only being attested in Hittite and Tocharian. The

problem, ultimately, is whether the number of such cognates is far greater within the proposed subgroup than across its borders. In the present state of knowledge, it is difficult to discern areal patterns in the witnesses of higher level (i.e. PMP) etyma and so we can only accept on good faith that such a preponderance of lexical evidence is not found in other areas and that the reconstructed etyma do not have cognates outside the relevant languages.[xiv] One potential heuristic for lexically based subgroups is the crispness of their borders. If we find a considerable amount of what Blust refers to as "leakage", then it is impossible to rule out that the entire distribution of certain etyma is due to millennia of "leakage", that is, horizontal spread. It is perhaps here that Proto-Philippines fares the worst.

Blust (2007) shows on the basis of linguistic evidence that the so-called "Sea Gypsies" of the Philippines, who speak languages belonging to the Sama-Bajaw subgroup (Pallesen 1985, Akamine 2005, Kaufman 2024), are relative newcomers to the Philippines and have origins in southern Borneo. Yet over 10% of the purported Proto-Philippine innovations are found in these languages along the southern border of the Philippines. Sometimes diagnostic sound changes indicate that the words are loans but many are phonologically ambiguous. Blust posits that even without diagnostic sound changes, all such words should be taken as loans because the chance of a higher level (i.e. PMP) etymon only surviving in a Sama-Bajaw language and an adjacent Philippine language is very small. But the large number of such words in Sama-Bajaw languages of the Philippines raises the more serious problem that the distribution of these words is entirely through horizontal transmission to begin with.[xv]

Lexical innovations have played a role in most subgrouping proposals in Austronesian but it appears that the subgroups whose sole basis is lexical have been the subject of frequent challenges and critiques for reasons that were first enumerated in the Austronesianist literature by Zorc (1982:313) and, as pointed out by Pereltsvaig and Martin (2015:70), have been understood since Meillet (1908). In view of this, Zorc (1986:155) proposes a classification of lexical innovations with varying evidential values and Smith (2017) proposes stricter conditions on lexical evidence with the four principles in (21),

(21) **Principle 1:** the innovation should be a replacement
Principle 2: the innovation should be robustly attested both in a number of individual languages (justifying their reconstruction to a protolanguage within the Philippine group) and in a number of microgroups (to justify their reconstruction to Proto-Philippines)
Principle 3: the sound correspondences between innovations must be regular
Principle 4: the innovations should be geographically noncontiguous

Important here is the idea that replacement innovations should hold more weight than non-replacement innovations but ultimately Blust's (2019, 2022) Proto-Philippine proposal demonstrates that even the strictest adherence to the principles in (21) and a virtual mountain of etyma do not insure general acceptance without support from other linguistic areas.

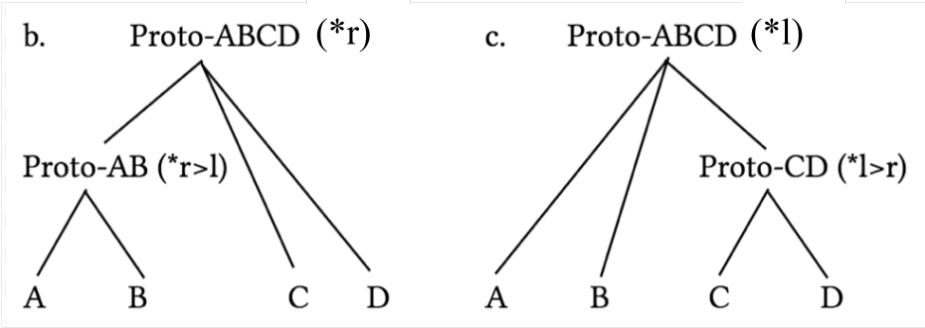
4. Subgrouping without the comparative method

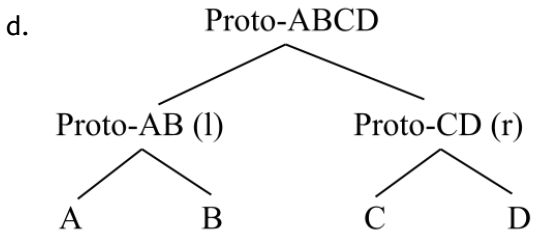
In this section, I review lexicostatistics together with what can be called probabilistic-computational phylogenetics (unfortunately, often referred to simply as "phylogenetic methods", a term which should not exclude the classic comparative method) with regard to Austronesian languages. These methods are treated more fully by Ceolin et al. (this volume) and McMahon and McMahon (2005). While statistical approaches differ, most are

crucially similar in setting aside the distinction between innovations and retentions which is the centerpiece of the comparative method. On this basis, it makes better sense to term such approaches 'resemblance based' or 'phenetic', following the practice in biological taxonomy, as they are based solely on the current state of languages without any predetermined theory of language change per se. While recent computational approaches can, in principle, take any feature of a language to evaluate possible family trees, some of the most prominent examples of this method, for both the Austronesian and Indo-European families, have employed lexical cognacy as the key character. Let us put this aside for the moment and assume a model that takes phonological aspects of cognates as the characters of interest, as in the Comparative Method. We can addend (1) above, repeated here as (22), with (d), to represent a phenetic method that subgroups solely according to shared similarities rather than shared innovations (but see Pellard et al, this volume, and Wichmann 2010 for other possibilities).

(22)a.

A	B	C	D
l	l	r	r





As discussed earlier, correspondences as in (22) would either yield a reconstruction *r, in which case the change *r>l becomes evidence for the Proto-AB subgroup, or *l, in which case the change *l>r becomes an argument for the Proto-CD subgroup. Most resemblance based models, on the other hand, would treat AB as a clade on par with CD, schematized by the tree in (24d), with retentions given the same strength as innovations.[xvi]

Lexicostatistics, a technique formalized by Morris Swadesh (1950 *et seq.*), involves measuring relatedness by counting cognates in a standard list of basic vocabulary. The linguist creates pairwise relations between all languages under evaluation, which then yields subgrouping hypotheses based on the number of shared etyma (Swadesh 1959). Unlike the comparative method, Lexicostatistics yields a continuum relation between all compared languages. If subgroups are imposed on this continuum, they are based on arbitrary percentages of lexical similarity.

It may be easily forgotten that for much of the latter part of the 20th century, lexicostatistics was an indispensable part of nearly all new subgrouping claims, including those works that were primarily based on the comparative method. Despite the obsolescence of lexicostatistics due to its inability to overcome problems of contact and borrowing and its dependence on regular rates of lexical replacement, the residue of this method is very much present in the most widely accepted present-day family trees of the

world's phyla (i.e. those found in Glottolog [Hammarström et al. 2024] and Ethnologue [Eberhard et al. 2023]), especially those which have not been subject to rigorous application of the comparative method.

Dyen (1963) undertook one of the most wide-ranging and rigorous applications of lexicostatistics in his pursuit of a family-wide classification of the Austronesian family and arrived at the conclusion that the Austronesian homeland was in the vicinity of Papua New Guinea. A southern homeland for Austronesian was (and remains) completely at odds with all else known about Austronesian history and is now recognized as erroneous. The error, first pointed out by Grace (1966), was clear; the mixing of Austronesian speaking newcomers with dozens if not hundreds of non-Austronesian groups on the northern New Guinea coast led to diversification of the lexicon through widespread borrowing and calquing. Thus, Dyen found that the center of lexical diversity was squarely in the Papuan region and, based on the principle of least moves (Sapir 1916, Dyen 1965), it was concluded that this area must be the Austronesian homeland. In contrast, the comparative method points strongly to Taiwan as the homeland, a hypothesis that has been since confirmed by archeology and genetic studies. The study did, however, succeed in exposing a fatal weakness of lexicostatistics and similar methods in the lack of their ability to differentiate shallow diversity (i.e. derived from recent contact) from deep diversity (i.e. that between primary branches of a family).

Resemblance based approaches have recently been resurrected on a large scale as part of probabilistic-computational phylogenetics. While the methods are now far more sophisticated and avoid some of the pitfalls of lexicostatistics (Greenhill and Gray 2009), current approaches are still restricted to comparing a standardized list of lexemes and generally disregard all that the field of historical linguistics has learned about directionality in language change. Consequently, despite important advances, computational phylogenetic methods fall prey to many of the same criticisms leveled against lexicostatistics (Donohue et al 2012a,b,c). Interestingly, the Austronesian tree deduced by Greenhill and Gray (2009) agrees in large part with subgrouping hypotheses derived via the comparative method, although for most linguists, there are too many unknowns in this method to treat it as a significant replication. For instance, the effect of "priors" (predetermined assumptions which the program uses to reduce possible trees) and the curation of the underlying database are not entirely clear.[xvii] Another cause for concern is the difficulty (if not outright impossibility) of replicating the results, as deriving the most probable subgroups for a large family like Austronesian is a process that requires enormous computing power and there are, by design, elements of chance in the process that would yield slightly different results even if the program was run with exactly the same variables.

Despite all the above, it should be emphasized that phenetic approaches are not expected to differ greatly from the comparative method in their subgrouping results, as most subgroups are contiguous, many maintain inter-group relations and lexical innovations can of course provide a real phylogenetic signal. It was for this reason that in the preceding case studies we focused on examples of non-contiguous relations that had been uncovered by the comparative method. When we zero in on where results from computational phylogenetic studies diverge from those of the comparative method, we find that it is precisely in these cases that computational phylogenetic methods fare the worst. For example, Donohue et al. (2012c) show that Gray et al. (2009) incorrectly place Samoan and Tongan in an exclusive subgroup based on surface similarities. The comparative method shows, however, that Samoan subgroups more closely with its neighbors to the east despite being obscured by contact relations. Similarly, recent work by King et al. (2023) on the Philippine languages concludes that geographical neighbor relations are the prime mover in Philippine phylogeny:

> "Our results show a dominant geographical signal. In our analysis, several groups considered to be only distantly related to other Philippine

> languages are strongly supported as sister groups to their geographic neighbors."

In support of this, they offer examples from several areas within the region:

> "Manide-Alabat, considered a deep branch of Philippines not closely related to any other (17), is grouped with neighboring Tagalog along with Sinauna. Conversely Kagayanen, a geographically isolated outlier of the Manobo micro-group, does not group with other Manobo languages."

> "The Northern Mangyan languages have been considered as relatives of the Central Luzon group (48, 49), but here we find strong support for a relationship with the Southern Mangyan languages, also from the island of Mindoro, in agreement with a lexicostatistical study (50)"

Rather than interpreting these results as a potential failure of method, the authors adduce it as evidence of the comparative method's failure, but the groupings they suggest are incorrect for clear reasons. Manide-Alabat and the language referred to as "Sinauna" are well known to have borrowed heavily from Tagalog (Lobel 2010, Santos 1975, Lobel & Surbano 2019). Similarly, Kagayanen is uncontroversially a Manobo language, as shown by Elkins (1974), but is well known to have borrowed heavily from Central Philippine languages, as Kagayanen speakers have been separated for many centuries from other Manobo groups on the remote Cagayancillo islands between Palawan and Negros, where Hiligaynon and Tagalog are the dominant lingua francas (Pebley and Payne 2024). Likewise, heavy contact between languages of the Northern Mangyan group and their neighbors on the southern half of the island has led to a number of isoglosses that cross subgroups. It is only through Zorc's (1974a) careful analysis of the native vocabulary and functional morphology of the North Mangyan languages that unusual correspondence sets emerge which set them apart clearly from their dominant Central Philippine neighbors to the south. The reason that King et al.'s (2023) results agree so well with earlier lexicostatistical studies is because the methods are, despite ample dissent, very similar in essence. In every case where computational phylogenetics diverges from the comparative method, it suggests that a language groups with neighboring languages with which it has had a long contact history, obscuring deeper connections. This replicates the more prominently discussed errors in computational phylogenetic approaches to Indo-European language history. As Pereltsvaig & Lewis (2015) point out, Bouckaert et al.'s (2012) phylogeny incorrectly groups Polish with its Eastern Slavic neighbors; Romanian as branching from Romance earlier than Sardinian; and Romani as a very early branch of Indo-Aryan due to its unique lexicon (which originates from later borrowings from non-Indo-Aryan languages). All of these errors are due to well known histories of contact. In no case have computational phylogenetic methods offered surprising connections between non-contiguous languages, as the comparative method has.[xviii]

It is well recognized that the family tree model of language, whose introduction is generally attributed to August Schleicher and Friedrich Schlegel, preceded similar models in biology, beginning with Darwin's *Origin of the Species*. What may come as more of a surprise to linguists is that the comparative method, with its focus on innovation, was almost a century ahead of its counterpart in biology. The so called "raging cladists" of evolutionary biology rediscovered the work of Hennig 1950 and proceeded to turn the received phenetic taxonomies upside down on the basis of strict adherence to innovations. Their most famous victim was the fish, as some species, like the lungfish, share crucial innovations with cows and thus subgroup more closely with bovines than with their ersatz aquatic companions. By eliminating retentions as a subgrouping heuristic, much unexpected progress was made in understanding the interrelations between living organisms, despite trenchant resistance from traditionalists.[xix]

Much of the computational phylogenetic literature gives the impression that biology has long moved on from innovation-based classification, cf. Atkinson & Gray (2005:520):

> "During the last 50 years, computational phylogenetic methods and statistical inference have revolutionized evolutionary biology. A burgeoning of sequence data has produced enormous databases that can only be investigated using computational techniques. Conversely, the field of linguistics, haunted perhaps by the "ghost of glottochronology past," has remained curiously averse to computational phylogenetic methods."

But dissenting voices in the field of Biology show that the picture is not nearly as neat. Brower (2020) advances criticisms that are entirely parallel to those of the skeptical linguists described above:

> "There is a certain Promethean arrogance manifest in the computational phylogeneticists' efforts to usher phenetics, taxonomic congruence, and other demonstrably defective methodologies back into the phylogenetic arena under the guise of statistical sophistication and algorithmically efficient heuristics. The window dressing may be more elaborate, but we've seen it all before. At their conceptual cores, most of these methods are neither novel nor useful to biological systematists."

Yet none of the above should lead us to ignore the very serious problem at the heart of subgrouping via the comparative method; namely, it is an inherently probabilistic enterprise with no clear way to calculate or express the probabilities involved and is furthermore highly vulnerable to cherry picking on the art of the linguist. Unfortunately, relatively little effort has been made in developing computational methods to calculate the probabilities of potentially subgroup-defining innovations using our accumulated knowledge of language change. A happy synthesis of the comparative method's transparency with robust statistical models to discern the direction of a change as well as the possibility of chance convergence and borrowing has the potential to advance subgrouping by leaps and bounds but this line of research has not yet attracted the same level of interest in computational circles as phenetic methods (but see Ringe et al. 2002 and references therein).[xx] A kindred problem identified by the computational phylogenetic literature is that the comparative method offers no simple way to quantify the strength of a subgrouping hypothesis. Haspelmath (2004b:216) notes that this has the unfortunate result of putting many poorly supported subgroups on par with secure ones. A real advantage of computational phylogenetic models is that they generate a precise rating for how well each subgroup fits the data, according to the algorithm of choice. Finally, it was noted that geography is often a silent partner in subgrouping decisions despite having no formal status in the comparative method. This is another area where improved computational approaches could remove researcher bias.

5. Conclusion

As noted earlier, in regions where our understanding of linguistic history is weak, neighboring languages spoken by communities in contact and sharing similar cultures are often subgrouped together with scanty evidence. Resources such as the Ethnologue (Eberhard et al. 2023) and Glottolog (Hammarström et al. 2024) necessarily draw upon sources using various methodologies with different confidence levels, although such discrepancies are not represented in the trees themselves. Hence the notion of a subgroup

in the field's premier databases remains essentially undefinable. Given this state of affairs, I conclude with four goals towards a better future for linguistic subgrouping:

- thorough isogloss mapping
- embracing conflict
- overcoming bias
- a more equitable marriage of computational techniques and the comparative method

Isogloss mapping, which has long been understood as the starting point of dialectology and was the end goal of many large scale linguistic projects around the world, is increasingly foregone in favor of skipping to the conclusion that a given number of features constitute subgroup defining innovations. The recent history of Austronesian subgrouping indicates that many innovations originally claimed of a particular protolanguage are in fact spread far and wide beyond the borders of the putative subgroup and must thus be reconsidered as evidence of common descent (cf. Donohue & Grimes 2008, Grimes & Edwards forthcoming, Smith 2017, Smith 2023). A return to thorough isogloss mapping for all components of the grammar and lexicon is a necessary precursor to establishing uniquely shared innovations.

In subgrouping via the comparative method, there exists an unfortunate temptation to sweep conflicts under the rug, perhaps a symptom of dendromania, the belief that language *classification* via a family tree is the only real goal of subgrouping. On the contrary, a loftier goal is to understand as much as possible about the history of a language and its speakers (even putting aside the grander goals of learning about language change and thus Language itself). Dyen (1956), Biggs (1965), Pallesen (1985) and Blust (1992) are classic examples in the Austronesian literature of how fruitful the exploration of irregular correspondences can be. In each case, apparent conflicts led to teasing apart multiple strata in the languages under study and ultimately distinguishing contact relations from original inheritance. Thus, finding and accounting for irregularity and conflict is just as meaningful as fitting a language in a tree, indeed it is a necessary step to proper classification, yet it depends on subgrouping via the comparative method as its starting point (cf. Jacques and List 2019).

Progress in subgrouping has been to a very large extent contingent on overcoming researcher biases of various sorts, ranging from the physical phenotype and culture of the speaker community, to geography, linguistic typology, social prestige and the existence of written traditions. As mentioned earlier, when confronted with the long written tradition of Old Javanese together with its complex, stratified society, it was assumed that the language reflected the ancient roots of the entire family better than the far less celebrated, unwritten tribal languages of the Philippines and Taiwan, and yet it is only by close examination of these latter languages that anything approaching Proto-Malayo-Polynesian and Proto-Austronesian can be reconstructed. Likewise, due to the phenotypical similarity of Fijians with non-Austronesian populations of Papua and Melanesia (the latter, a European-devised region made on the basis of the population's skin color), it was assumed that Fijian was a mixed (part Austronesian, part "Papuan") language. It was only by unbiased application of the comparative method that Fijian was shown to represent an early branch of Proto-Oceanic, far more conservative than other languages of the region spoken by groups that better fit a Southeast Asian phenotype. Similar examples could be multiplied (e.g. Inati of the Central Philippines and Malagasy, both of which are surprisingly conservative given the significant non-Austronesian component in the speaker population). The case studies reviewed in §3 also illustrate clearly that general typology tends to obscure rather than clarify the internal structure of language families.

Finally, computational methods have been brought to bear on subgrouping in dramatic fashion but much of this work has bypassed critical elements of the comparative

method in favor of Bayesian methods popular in biology. While it is agreed upon that the application of the comparative method tacitly relies on probability, there should be more effort dedicated to modelling these probabilities while still making use of the core insight of the comparative method, namely, that innovations, especially those that are linguistically unusual and resistant to borrowing, are key to deducing common descent. Work in the vein of Ringe et al. (2002) and Nakhleh et al. (2005), which applies computational methods to finding the most likely subgrouping on the basis of phonological and morphological innovations, demonstrates that this is a fruitful path forward.

Related Articles (See Also)

Article ID
WAVE MODEL
MERGERS and SPLITS
FAMILY TREE MODEL
COMPUTATIONAL PHYLOGENETIC MODELS
COMPARATIVE METHOD AND COMPARATIVE RECONSTRUCTION
CONVERGENCE AND LINGUISTIC AREAS
CONTACT AND BORROWING

References

Adelaar, K. Alexander. 1989. Malay influence on Malagasy: linguistic and culture-historical implications. Oceanic Linguistics 28: 1–46.

Adelaar, K. Alexander. 1995. Asian roots of the Malagasy: a linguistic perspective. Bijdragen tot de Taal-, Land-, en Volkenkunde 151: 325–356.

Adelaar, K. Alexander. 2005. Malayo-Sumbawan. Oceanic Linguistics 44(2): 357–388.

Adelaar, K. Alexander. 2010. The amalgamation of Malagasy. In John Bowden, Nikolaus P. Himmelmann, and Malcolm D. Ross (eds.), A journey through Austronesian and Papuan linguistic and cultural space, papers in honour of Andrew K. Pawley. 161–178. Canberra: Pacific Linguistics.

Adelaar, K. Alexander. 2012. Malagasy phonological history and Bantu influence. Oceanic Linguistics 51: 123–159.

Adelaar, K. Alexander. 2021. South Borneo as an ancient Sprachbund area. *Wacana* 22(1):81-101.

Aikhenvald, Alexandra and Robert Dixon. 2001. Areal Diffusion and Genetic Inheritance: Problems in Comparative Linguistics. Oxford, New York: Oxford University Press.

Akamine, Jun. 2005. ‘Sama (Bajau)’. In K. Alexander Adelaar and Nikolaus P. Himmelmann (eds.), The Austronesian languages of Asia and Madagascar. London: Routledge, 377- 396.
Aldridge, Edith. 2010. Directionality in word-order change in Austronesian languages. In Anne Breitbarth, Christopher Lucas, Shelia Watts, David Willis (eds) Continuity and change in grammar. Amsterdam/Philadelphia: John Benjamins, 169–180.

Aldridge, Edith. 2016. Ergativity from Subjunctive in Austronesian Languages. Language and Linguistics 17:27-62.

Aldridge, Edith and Yuko Yanagida. 2021. Two types of alignment change in nominalizations: Austronesian and Japanese. Diachronica 38(3). 314-357.

Atkinson, Quentin D. and Russel D. Gray. 2005. Curious parallels and curious connections: Phylogenetic thinking in Biology and Historical Linguistics. Systematic Biology 54(4): 513-526.

Baklanova, Ekaterina and Kate Bellamy. 2023. Spanish Suffixes in Tagalog: The Case of Common Nouns. In Marian Klamer and Francesca R. Moro (eds.), Traces of Contact in the Lexicon, 307–347. Amsterdam: Brill.

Beck, David. 2023. Morphological diffusion and the internal sub-grouping of Central Totonac. Language Dynamics and Change 13(3), 1–52.

Bellwood, Peter. 2017. First Islanders: Prehistory and Human Migration in Island Southeast Asia. New Jersey: Wiley-Blackwell.

Biggs, Bruce G. 1965. Direct and indirect inheritance in Rotuman. Lingua 14: 383–415.

Blevins, Juliette and Andrew Garrett. 1998. ‘The origins of consonant-vowel metathesis.’ Language 74, 508–556.

Blood, Doris W. 1978. ’Some aspects of Cham discourse structure.’ Anthropological Linguistics 20.3:110-132.

Blust, Robert A. 1977. The Proto-Austronesian pronouns and Austronesian subgrouping: a preliminary report. University of Hawai’I working papers in Linguistics 9(2): 1-15.

Blust, Robert A. 1991. The Greater Central Philippines hypothesis. Oceanic Linguistics 30:73–129.

Blust, Robert A. 1992. On speech strata in Tiruray. In Malcolm D. Ross, ed., Papers in Austronesian Linguistics, No. 2:1-52. Pacific Linguistics A82.

Blust, Robert A. 1995. The position of the Formosan languages: Method and theory in Austronesian comparative linguistics. In Austronesian studies relating to Taiwan, ed. by Paul Jenkuei Li, Cheng-hwa Tsang, Ying-kuei Huang, Dah-an Ho, and Chiu-yu Tseng, 585– 650. Symposium Series of the Institute of History and Philology, Academia Sinica, No. 3. Taipei: Academia Sinica

Blust, Robert A. 1998.The position of the languages of Sabah. In Ma. Lourdes S. Bautista, ed., Pagtanáw: Essays on language in honor of Teodoro A. Llamzon: 29-52. Manila: Linguistic Society of the Philippines.

Blust, Robert A. 1999. “Subgrouping, circularity, and extinction: Some issues in Austronesian comparative history”. In: Selected Papers from the Eighth International Conference on Austronesian Linguistics, Elizabeth Zeitoun and Paul Jen-kuei Li (eds.), pp.31-94. Taipei: Academica Sinica.

Blust, Robert A. 2005. The linguistic macrohistory of the Philippines: Some speculations. In Current issues in Philippine linguistics and anthropology: Parangal kay Lawrence A. Reid, ed. by Hsiu-chuan Liao and Carl R. Galves Rubino, 31–68. Manila: Linguistic Society of the Philippines and SIL International.

Blust, Robert A. 2007. The Linguistic position of Sama-Bajaw. Studies in Philippine Languages and Cultures 15: 73–114.

Blust, Robert A. 2010. The Greater North Borneo hypothesis. Oceanic Linguistics 49:44–118

Blust, Robert A. 2013. The Austronesian languages, 2nd ed. Canberra: Pacific Linguistics.

Blust, Robert A. 2019. The resurrection of Proto-Philippines. Oceanic Linguistics 58(2):153-256.

Blust, Robert. 2020. Response to Comments on "The Resurrection of Proto-Philippine" Oceanic Linguistics 59(1/2): 450-479.

Blust, Robert. 2022. Proto-Philippine Addenda: Theory, Method and Data. Oceanic Linguistics 61(1): 322-404.

Blust, Robert and Victoria Chen. 2017. The pitfalls of negative evidence 'Nuclear Austronesian', 'Ergative Austronesian', and their progeny. Language and Linguistics 18(4): 577-621.

Bopp, F., 1816, Über das Conjugationssystem der Sanskritsprache in Vergleichung mit jenem der griechischen, lateinischen, persischen und germanischen Sprache. Nebst Episoden aus dem Ramajana und Mahabharata in genauen metrischen Übersetzungen aus dem Originaltexte und einigen Abschnitten aus den Vedas. Heraus- gegeben und mit Vorerinnerungen begleitet von Dr. K. J. Windischmann, Frankfurt/Mandreäsche Buchhandlung.

Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson. 2012. Mapping the Origins and Expansion of the Indo-European Language Family. Science 337:957-960.DOI:10.1126/science.1219669

Brower, Andrew V. Z. 2020. No background in biology is assumed. Cladistics 36(4):437-442.

Brugman, Karl. 1884. Zur frage nach den Verwandtschaftsverhältnissen der indogermanschen Sprachen. Separatabdruck Internationale Zeitschrift für allg. Sprachwissenschaft 1(253). Leipzig.

Campbell, Lyle and Alice Harris. 1995. Historical Syntax in Cross-Linguistic Perspective. Cambridge: CUP.

Charles, Mathew. 1974."Problems in the Reconstruction of Proto-Philippine Phonology and the Subgrouping of the Philippine Language". Oceanic Linguistics 13(1/2): 457–509.

Chen, Victoria. 2017. A reexamination of the Philippine-type voice system and its implications for Austronesian primary-level subgrouping. Ph.D diss. University of Hawai'i at Mānoa.

Chen, Victoria , Kristina Gallego, Jonathan Kuo, Isaac Stead, and Benjamin van der Voorn. 2024. Contact or inheritance? New evidence on the Proto-Philippines debate. ms.

Chen, Victoria Jonathan Kuo, Maria Kristina Gallego, and Isaac Stead. 2022. "Is Malayo-Polynesian a primary branch of Austronesian? A view from morphosyntax." *Diachronica* 39(4):449-489.

Cumming, Susanna. 1991. Functional Change: the Case of Malay Constituent Order. Berlin: Walter de Gruyter.

Dahl, Otto Christian. 1951. Malgache et Maanjan. Oslo: Arne Gimnes Verlag.

Dahl, Otto Christian. 1973. Proto-Austronesian. Scandinavian Institute of Asian Studies Monograph Series, No. 15. Lund, Studentlitteratur.

Dempwolff, Otto. 1934. Vergleichende Lautlehre des austronesischen Wortschatzes, Band 1: Induktiver Aufbau einer indonesischen Ursprache. Beihefte zur Zeitschrift für Eingeborenen-Sprachen 15. Berlin: Dietrich Reimer.

Dempwolff, Otto. 1937. Vergleichende Lautlehre des austronesischen Wortschatzes, Band 2: Deduktive Anwendung des Urindonesischen auf austronesische Einzelsprachen. Beihefte zur Zeitschrift für Eingeborenen-Sprachen 17. Berlin: Dietrich Reimer.

Dixon, R.M.W. 1997. The rise and fall of languages. Cambridge: Cambridge University Press

Donohue, Mark. 2007. Word order in Austronesian: from north to south and west to east. Linguistic Typology 11 (2): 351-393.

Donohue, Mark, and Charles E. Grimes. 2008. Yet more on the position of the languages of eastern Indonesia and East Timor. Oceanic Linguistics 47 (1): 115-159.

Donohue, Mark, Tim Denham and Stephen Oppenheimer. 2012a. Uncoupling inheritance and diffusion: a lexical-based methodology detects social distance. Diachronica 29 (4): 502-522.

Donohue, Mark, Tim Denham and Stephen Oppenheimer. 2012b. Consensus and the lexicon in historical linguistics. Diachronica 29 (4): 538-546.

Donohue, Mark, Tim Denham, Stephen James Oppenheimer. 2012c. New methodologies for historical linguistics? Calibrating a lexicon-based methodology for diffusion vs. subgrouping. Diachronica 29:505-522

Donohue, Mark, Simon Musgrave, Bronwen Whitting, and Søren Wichmann. 2011. Typological feature analysis models linguistic geography. Language 87: 369–383.

Donohue, Mark, Søren Wichmann and Mihai Albu. 2008. Typology, areality and diffusion. Oceanic Linguistics 47 (1): 223-232.

Dyen, Isidore. 1953. Reviewed of Malgache et Maanjan: Une comparaison linguistique by Otto Chr. Dahl. Language 29(4):577-590.

Dyen, Isidore. 1956. The Ngaju-Dayak 'old speech stratum'. Language 32: 83–87.

Dyen, Isidore. 1965. "Language distribution and migration theory". Language. 32 (4): 611–626.

Edwards, Owen. 2021. Rote-Meto Comparative Dictionary. Asia-Pacific Linguistics Series. Canberra: ANU Press.

Grimes, Charles and Owen Edwards. forthcoming. Austronesian languages of Eastern Indonesia. Canberra: ANU Press.

Eberhard, David, M., Gary F. Simons, Chuck D. Fennig. 2023. Ethnologue: Languages of the World (26th ed.). SIL International.

François, Alexandre. 2014. “Trees, Waves and Linkages: Models of Language Diversification”, in Claire Bowern and Bethwyn Evans (eds.), The Routledge Handbook of Historical Linguistics, pp. 161–189. London: Routledge.

Givon, Talmy. 1979. On Understanding Grammar. New York: Academic Press.

Grace, George. 1966. Austronesian lexicostatistical classification: A review article. Oceanic Linguistics 5(1): 13–31.

Greenberg, Joseph H. 1957. "The problem of linguistic subgroupings", in Essays in Linguistics. Chicago: University of Chicago Press.

Greenhill, Simon and Russell D. Gray. 2009. Austronesian language phylogenies: myths and misconceptions about Bayesian computational methods. In Alexander Adelaar and Andrew Pawley (eds.), Austronesian historical linguistics and culture history: a festschrift for Robert Blust, pp. 375-398. Canberra: Pacific Linguistics.

Hammarström, Harald, Robert Forkel, Martin Haspelmath, Sebastian Bank. 2024. Glottolog 5.0. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at https://glottolog.org)

Harrison, Shelly. 2003. On the Limits of the Comparative Method. In B. D. Joseph, & R. D. Janda (Eds.), The Handbook of Historical Linguistics, pp. 213-242. Oxford: Blackwell.

Haspelmath, Martin. 2004. How hopeless is genealogical linguistics, and how advanced is areal linguistics? Studies in Language 28(1):209-223.

Haspelmath, Martin & Uri Tadmor (eds.). 2009. Loanwords in the world’s languages: a comparative handbook. Berlin: Mouton de Gruyter.

Haugen, Einar, 1950. The analysis of linguistic borrowing. Language 26, 210–231.
Heine, Bernd and Tania Kuteva. 2002. World Lexicon of Grammaticalization. Cambridge: Cambridge University Press.

Hennig, Willi. 1950. Grundzüge einer Theorie der phylogenetischen Systematik, Berlin: Deutscher Zentralverlag.

Himmelmann, Nikolaus P. 2005. The Austronesian Languages of Asia and Madagascar: Typological Charactersistics. In A. Adelaar and N. P. Himmelmann (eds.) The Austronesian Languages of Asia and Madagascar, 110-181. London: Routledge.

Hopper, Paul J. and Elizabeth Traugott. 2003. Grammaticalization. Cambridge: Cambridge University Press.

Hudson, Alfred B. 1967. The Barito Isolects of Borneo: A Classification Based on Comparative Reconstruction and Lexicostatistics. Ithaca: Cornell University Press.

Jacobson, Marc. 1999. Inabaknon wordlist. Available at:
https://www.trussel2.com/acd/acdinab-a.htm

Jacques, Guillaume and Johann-Mattis List. 2019. Save the trees: Why we need tree models in historical linguistics (and when we should apply them). Journal of Historical Linguistics 9(1):128-166.

Kaufman, Daniel. 2009a. South Sulawesi pronominal clitics: form, function and position. Studies in Philippine Languages and Cultures 17, pp.13-65.

Kaufman, Daniel. 2009b. “Austronesian typology and the nominalist hypothesis”. In Alexander Adelaar & Andrew Pawley (eds.), Austronesian Historical Linguistics and Culture History: A Festschrift for Robert Blust. pp.179-226.

Kaufman, Daniel. 2014. “The syntax of Indonesian imposters”. In Chris Collins (ed.), Cross-linguistic Studies of Imposters and Pronominal Agreement, pp. 89-120. Oxford: Oxford University Press.

Kaufman, Daniel. 2024. The Sama-Bajaw languages. In Antoinette Schapper and Sander Adelaar (eds.) Oxford Guide to the Malayo-Polynesian languages of Southeast Asia. Oxford: Oxford University Press.

King, Benedict, Simon J. Greenhill, Lawrence A. Reid, Malcolm Ross, Mary Walworth, and Russell Gray. 2023. Bayesian Phylogenetic Analysis of Philippine Languages Supports a Rapid Migration of Malayo-polynesian Languages. SocArXiv. March 31. doi:10.31235/osf.io/re8m6.

Liao, Hsiu-Chuan. 2020. A reply to Blust (2019) The resurrection of Proto-Philippines. Oceanic Linguistics 59(1/2):426-449.

Lobel, Jason W. 2010. Manide: an undescribed Philippine language. Oceanic Linguistics 49(2). 480-512.

Lobel, Jason William & Orlando Vertudez Surbano. 2019. Notes from the Field: Remontado (Hatang-Kayi): A Moribund Language of the Philippines. Language Documentation & Conservation 13. 1-35.

Lynch, John, Malcolm Ross and Terry Crowley. 2002. The Oceanic Languages. London: Routledge.

Mahdi, Waruno. 1986. Morphophonologische Besonderheiten und historische Phonologie des Malagasy. Berlin: Dietrich Reimer.

J.P. Mallory and Douglas Q. Adams. 2006. The Oxford Introduction to Proto-Indo-European and the Proto-Indo-European World. Oxford: Oxford University Press.
Matras, Yaron. 2007. Language Contact. Cambridge: Cambridge University Press.

McMahon, April and McMahon, Robert. 2005. Language Classification by Numbers. Oxford University Press.

Mead, David. 2002. Proto-Celebic focus revisited. In Wouk, Fay, and Malcolm Ross (eds.) The history and typology of western Austronesian voice systems. Pacific Linguistics 518, pp. 143–77. Canberra: Australian National University.

Meillet, Antoine.1908. Les dialectes indo-européens. Paris: H. Champion.

Miller, Lulu. 2020. Why Fish Don’t Exist. New York: Simon & Schuster.

Mills, Roger Frederick. 1975. Proto South Sulawesi and Proto Austronesian Phonology. Ph.D. diss. University of Michigan.

Muysken, Pieter. 1981. Halfway between Quechua and Spanish: The case for relexification. In Highfield, Arnold and Valdman, Albert (eds.), Historicity and variation in creole studies, 52-78. Ann Arbor: Karoma.

Nakhleh, Luay, Don Ringe, and Tandy Warnow. 2005. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. Language 81: 382–420.

Næss, Åshild. 2013. From Austronesian voice to Oceanic transitivity: Äiwoo as the 'missing link'. Oceanic Linguistics 52(1): 106-124.

Næss, Åshild. 2021. Voice and Valency Morphology in Äiwoo. Oceanic Linguistics 60(1): 160-198.

Nichols, Johanna. 1992. Linguistic Diversity in Space and Time. Chicago: University of Chicago Press.

Norde, Muriel. 2009. Degrammaticalization. Oxford: Oxford University Press.

Pallesen, A. Kemp. 1985. Culture contact and language convergence. Philippine journal of linguistics: special monograph issue, 24. Manila: Linguistic Society of the Philippines.

Paz, Consuelo J. 1981. A reconstruction of Proto-Philippine phonemes and morphemes. Publication 3 of the Cecilio Lopez Archives of Philippine Languages and the Philippine Linguistics Circle. Diliman, QC: University of the Philippines.

Pebley, Carol J. and Thomas E. Payne. 2024. A Grammar of Kagayanen. Berlin: Language Science Press. DOI: 10.5281/zenodo.12755278

Pereltsvaig, Asya and Martin W. Lewis. 2015. The Indo-European Controversy: Facts and Fallacies in Historical Linguistics. Cambridge: Cambridge University Press.

Reid, Lawrence. 1982. The demise of Proto-Philippines. In Amram Halim, Lois Carrington, and S.A. Wurm (eds.), Papers from the Third International Conference on Austronesian Linguistics, vol.2: Tracking the travellers, 201-216. Canberra: Pacific Linguistics.

Reid, Lawrence A. 2020. Response to Blust "The resurrection of Proto-Philippines". Oceanic Linguistics 59(1/2):374-393.

Ringe, Don, Tandy Warnow, and Ann Taylor. 2002. Indo-European and computational cladistics. Transactions of the Philological Society 100: 59–129

Ross, Malcolm. 1988. Proto Oceanic and the Austronesian languages of western Melanesia. Canberra: Pacific Linguistics.

Ross, Malcolm. 2001. "Is there an East Papuan phylum? Evidence from pronouns." In The Boy from Bundaberg: Studies in Melanesian Linguistics in Honour of Tom Dutton, A. Pawley, M. Ross and D. Tryon (eds.). p. 301-321. 1st ed. Canberra: Pacific Linguistics.

Ross, Malcolm D. 2001b. "Contact-induced change in Oceanic languages in North-West Melanesia". In Alexandra Y. Aikhenvald; R. M. W. Dixon (eds.). Areal diffusion and genetic inheritance. Oxford: Oxford University Press. pp. 134–166.

Lynch, John, Malcolm Ross and Terry Crowley (eds.). 2002. The Oceanic Languages. Richmond, Surrey: Curzon Press.

Ross, Malcolm. 2005. "Pronouns as a preliminary diagnostic for grouping Papuan languages. In Papuan pasts: Cultural, linguistic and biological histories of Papuan-speaking peoples, Andrew Pawley, Robert Attenborough, Jack Golson and Robin Hide (eds.). 1st ed. p. 15-66. Canberra Australia: Pacific Linguistics.

Ross, Malcolm. 2008. “The integrity of the Austronesian language family: from Taiwan to Oceania”. In: Past Human Migrations in East Asia: Matching archaeology, linguistics and genetics. Alicia Sanchez-Mazas, Roger Blench, Malcolm D. Ross, Ilia Peiros and Marie Lin (eds.), p. 161-181. Great Britain: Routledge, Taylor & Francis Group.

Ross, Malcolm. 2009. “Proto-Austronesian verbal morphology: A reappraisal”. In Austronesian Historical Linguistics and Culture History: A festschrift for Robert Blust, Alexander Adelaar and Andrew Pawley (eds.), pp. 295–326. Canberra: Pacific Linguistics.

Ross, Malcolm. 2012. Just how different was Proto Oceanic from Proto-Malayo-Polynesian? Paper presented at the 12th International Conference on Austronesian Linguistics (12ICAL), Denpasar, Indonesia, July.

Ross, Malcolm. 2020. Comment on Blust “The resurrection of Proto-Philippines”. Oceanic Linguistics 59(1/2):366-373.

Ross, Malcolm, Andrew Pawley and Meredith Osmond (eds.). 1998. The lexicon of Proto Oceanic, the culture and environment of ancestral Oceanic society, 1, Material culture. Canberra: Pacific Linguistics.

Sagart, Laurent. 2004. The higher phylogeny of Austronesian and the position of Tai-Kadai. Oceanic Linguistics 43, 2: 411-444.

Santos, Pilar C. 1975. Sinauna Tagalog: A genetic study examining its relation with other Philippine languages. M.A. thesis, Ateneo de Manila.

Sapir, Edward. 1916. "Time Perspective in Aboriginal American culture: A Study in Method”. Geological Survey of Canada. Memoir 90. No. 13 Anthropological Series.

Sapir, Edward. 1921. Language: An introduction to the study of speech. New York: Harcourt, Brace & World.

Schleicher, August. 1853. Die ersten Spaltungen des indogermanischen Urvolkes. Allgemeine Monatsschrift für Wissenschaft und Literatur, 3:786-787.

Schmidt, Johannes. 1872. Die Verwandtschaftsverhältnisse der indogermanischen Sprachen. Weimar: H. Böhlau.

Schmidt, Wilhelm. 1906. Die Mon-Khmer-Valker, ein Bindeglied zwischen Valkern Zentralasiens und Austronesiens. Arch. Anthrop., Braunschweig, 5:59-109.

Sebeok, Thomas A. 1942. An examination of the Austroasiatic language family. Language 18: 206-217.

Seifart, Frank. 2020. AfBo: A world-wide survey of affix borrowing. Leipzig: Max Planck Institute for Evolutionary Anthropology. DOI: 10.5281/zenodo.3610155

Smith, Alexander D. 2017. The Western Malayo-Polynesian problem. Oceanic Linguistics 56(2): 435–489.

Smith, Alexander D. 2018. The Barito Linkage Hypothesis with a note on the position of Basap. *Journal of the Southeast Asian Linguistics Society* 11(1): 13–34.

Smith, Alexander D. 2023. Evidence and Models of Linguistic Relations: Subgroups, Linkages, Lexical Innovations, and Borneo. Oceanic Linguistics 62(2): 324-365.

Starosta, Stanley, Andrew K. Pawley & Lawrence A. Reid. 1982. The evolution of focus in Austronesian. In Amran Halim, Lois Carrington & S.A. Wurm (eds.), Papers from the third International Conference on Austronesian Linguistics. Vol. 2: Tracking the travellers (Pacific Linguistics C-65). Canberra: Research School of Pacific and Asian Studies, Australian National University, 145–170.

Swadesh, Morris. 1950. Salish internal relationships. International Journal of American Linguistics. 16 (4): 157–167.

Swadesh, Morris. 1959. Linguistics as an Instrument of Prehistory. Southwestern Journal of Anthropology. 15: 20–35.

Tadmor, Uri, Martin Haspelmath, and Bradley Taylor. 2010. Borrowability and the notion of basic vocabulary. Diachronica 27: 226–246.

Thurgood, Graham, Ela Thurgood, Li Fengxiang. 2015. A Grammatical Sketch of Hainan Cham: History, Contact, and Phonology. Pacific Linguistics vol. 643. Berlin: De Gruyter Mouton.

Thomason, Sarah Grey and Terrence Kaufman. 1988. Language Contact, Creolization, and Genetic Linguistics. Berkeley: University of California Press.

Thomason, Sarah G. and Daniel L. Everett. 2005. Pronoun borrowing. Berkeley Linguistic Society 27. 301-315.

van der Tuuk, H. N. 1865. Outlines of a Grammar of the Malagasy Language. Journal of the Royal Asiatic Society of Great Britain and Ireland.

Thurgood, Graham. 1999. From Ancient Cham to Modern Dialects: Two Thousand Years of Language Contact and Change. Honolulu: University of Hawai'i Press.

Thurgood, Graham. 2002. Phan Rang Cham. In K. Alexander Adelaar and Nikolaus Himmelmann (eds.), The Austronesian Languages of Asia and Madagascar, pp.489-512. London: Routledge Press.

Van den Berg, René. 1996. The demise of focus and the spread of conjugated verbs in Sulawesi. In Hein Steinhauer (ed.), Papers in Austronesian Linguistics 3:89–114. Pacific Linguistics A 84.

Von Humboldt, Wilhelm. 1836. Über die Kawi sprache auf der Insel Jawa. Berlin: Königlichen Akademie der Wissenschaften.

Weinrich, Uriel. 1968. Languages in Contact: Findings and Problems. Amsterdam: Mouton.

Wichmann, Søren. 2010. Internal language classification. In Continuum Companion to Historical Linguistics, Silvia Luraghi and Vit Bubenik (eds.), pp. 70-88. London: Coninuum Press.

Wolff, John. 1996. "The development of the passive verb with pronominal prefix in Western Austronesian languages", in Reconstruction, Classification, description: festschrift in honor of Isidore Dyen, Bernd Nothofer (ed.), pp. 15-40. Hamburg: Abera.

Yoon, Carol Kaesuk. 2009. Naming Nature: The Clash Between Instinct and Science. New York: W. W. Norton & Company

Zorc, David. 1974a. Internal and external relationships of the Mangyan languages. Oceanic Linguistics 13(1/2):561-600.

Zorc, David. 1974b. Towards a definitive Philippine wordlist - the qualitative use of vocabulary in identifying and classifying languages. Oceanic Linguistics 13:409-455.

Zorc, David. 1977. The Bisayan Dialects of the Philippines: Subgrouping and Reconstruction. Pacific Linguistics C.44. Canberra: The Australian National University.

Zorc, David. 1982. "Micro and macro grouping criteria, problems and procedures". In GAVA': Studies in Austronesian languages and cultures dedicated to Hans Kähler, Rainer Carle et al. (eds.) Band 17:305-320. Berlin: Dietrich Reimer Verlag.

Zorc, David. 1986. "The genetic relationships of Philippine languages". In Paul Geraghty, Lois Carrington, and S.A. Wurm (eds.) FOCAL II: Papers from the Fourth International Conference on Austronesian Linguistics, Series C-94, pp.147-173. Canberra: Pacific Linguistics.

Figure captions

Figure 1. A portion of the Rote-Meto family tree with areal features overlaid on a family tree (Edwards 2021)
Figure 2. Austronesian family tree (adapted from Ross 2008)

Tables

[Please insert any tables here]

ENDNOTES

[i] It has been conjectured that the characteristic innovations of certain Indo-European subgroups arose through contact with pre-Indo-European populations but this remains pure conjecture because so little is known about those languages. Thus, while we cannot rule out a major role for contact and areal phenomena in the diversification of the Indo-European family, it is far more difficult to demonstrate.

[ii] In the case of Tagalog, we can see the borrowing of the Spanish agentive suffixes attaching to native stems, as in (i)-(iii), (cf. Thomason and Kaufman 1988, Baklanova and Bellamy 2023).

(i)	laseng-go	(ii)	bulakbol-ero	(iii)	bungang-era
	drunk-AGT		wander-AGT.msc		mouth-AGT.fem
	'alcoholic'		'truant, vagrant'		'bigmouth, gossip'

[iii] One example of a paradigmatic change with probative value is the loss of third person plural forms throughout the South Sulawesi languages, delineating a border with most of their neighbors. A third person plural was reinnovated in several members of the subgroup (Kaufman 2009, Mills 1975), but never with a reflex of the original PMP form. However, a similar loss of third person plural also occurred in the Wotu-Wolio languages and several other languages adjacent to the South-Sulawesi subgroup, which indicates the potential of such a change to spread horizontally. The loss of the clusivity distinction in the first person plural is even poorer as a phylogenetic signal in this region, as it crosscuts several well-established subgroups, including languages of the South Sulawesi subgroup, the Muna-Buton subgroup, and Malayic, among others, and therefore has very little value for subgrouping, despite the purported stability of this feature globally. The diachronic stability of the clusivity distinction as well as plurality neutralization (which the loss of the third person plural may be classified as) is argued for at length by Nichols (1992) on a global scale.

[iv] Blust (1999) proposes nine Formosan subgroups as direct daughters of PAn based exclusively on sound change. More recent investigations based on morphosyntax (Aldridge 2016 and Ross 2009) and the lexicon (Sagart 2004) attempt to posit higher level subgrouping of Formosan languages.

[v] Interestingly, this metathesis did not apply regularly across the entire PMP lexicon. Some words, like PAn *tuqaS 'old' and *liseqeS 'nit, egg of a louse' yielded doublets in PMP, with one metathesized member and one unmetathesized member, possibly due to phonological conditioning, as several exceptions involve *qVS. Despite this, the metathesis has been unanimously viewed as very strong evidence.

[vi] This type of gestalt argument is made explicit by Adelaar (2005) in his proposal of a Malayo-Sumbawan subgroup. While recognizing the low probative value of the individual pieces evidence (because of their cross-linguistic commonality), he argues they gain value when shared by a group of languages in tandem. The subgroup has nonetheless had its critics (Blust 2010).

[vii] A difficulty for reconstructing Proto-Oceanic grammar is that many languages of what are now considered to be primary subgroups (e.g. Temotu, Admiralities, Yapese) are incompletely described or completely undescribed. Recent work examining some of the primary branch outliers (Ross 2012, Næss 2013, 2021) suggests that Proto-Oceanic grammar may have been more similar to PMP than previously thought.

[viii] In an interesting twist, Smith (2018) casts doubt on the unity of the Barito subgroup itself, arguing that it is not a proper subgroup but rather an innovation-defined linkage, where sound changes overlap in a stepwise fashion from one end of the Barito territory to the other. Despite this new understanding, Malagasy's connection to the Barito languages remains unchallenged.

[ix] An example of extreme typological convergence with non-Austronesian languages is well illustrated in Thurgood et al's (2014) description of Hainan Cham.

[x] The term Philippine language and Philippine-type language have been used in different ways in the Austronesianist literature (Himmelmann 2005), but it is clear that the geographic boundaries of the Philippines happens to correlate roughly with a typological group, whose members also extend into Sulawesi and parts of northern Borneo. The typological similarities in these languages (e.g. a four-way voice system, predicate-initial word order, case-marking proclitic determiners, ergative characteristics with an ergative-genitive syncretism, second-position pronominal and adverbial clitics) are all retentions from PMP rather than innovations and thus offer no support for a putative Proto-Philippine language. Rather, it is the loss of all these features in many areas south of the Philippines that requires explanation, either through independent innovations or contact induced change.

[xi] An immediate problem with this approach is that, under current conditions, the work of whittling would require years of intensive manual labor spent poring over an enormous number of lexical resources. Few scholars may be interested in dedicating such effort to verifying the distribution of each word in this list, although some of this work has already begun (Liao 2020, Zorc 2020).

[xii] Blust attempts to guard against this possibility by excluding words that could have a Tagalog loan distribution, words that refer to trade items, words that are only attested in a small number of contiguous Philippine subgroups, and words that show irregular phonological correspondences. But even these safeguards may be insufficient. First, Tagalog has only become the dominant language of the region in the last several hundred years. There is good evidence from loan words in every corner of the Philippines that multilingualism and contact played a long role in the development of all Philippine languages. Blust (1992), a pioneering work in this area, showed how the majority of the documented Tiruray lexicon has an external origin and how several strata, each with its own sound correspondences, can be teased apart. While Tiruray may present an extreme example, the existence of multiple strata in the lexicon is a widespread phenomenon and the lending languages are most often not the prestige languages of today. Thus, exclusion of words with Tagalog cognates can only guard against the most recent layer of borrowings.

[xiii] Calculating chance is notoriously difficult and depends on the number of phonemes in the words compared, the size of the relevant phonological inventories, the degree of semantic matching, and possibly other factors, as well.

[xiv] There is no comprehensive, unified lexical resource for Austronesian languages outside the reconstructions found in the ACD as there is for Austroasiatic languages in the form of the Mon-Khmer Etymological Dictionary (http://sealang.net/monkhmer/database/).

[xv] Blust (2019) writes: "It has also been shown that the Philippine-type languages of Sabah almost certainly are not part of the proposed Philippine subgroup, but have absorbed many loanwords as a result of the GCP

expansion into northern Sulawesi and northern Borneo (Blust 1998, 2010). Given that the Sama–Bajaw and Sabahan languages are not members of the Philippine group, it is notable that a substantial number of lexical comparisons that are otherwise restricted to Philippine languages are also found in one or both of them." This betrays a Norman Conquest model of borrowing, whereby words spread through rapid, likely militaristic, conquering events, but there is nothing at all in the archeological record nor the linguistic record of the region that suggests one off conquering events were a primary means of lexical diffusion. Rather, ancient patterns of raiding and intermarriage with neighboring groups appear to have taken place since the earliest times and were the rule throughout the region rather than the exception.

[xvi] The occasional claim (cf. Greenhill and Gray 2012:525) that the comparative method itself does not distinguish innovations from retentions but rather induces innovations from subgrouping hypotheses is backwards. There is a well-known potential for circularity but the study of language change provides strong independent evidence for the directionality of many changes. It is thus the subgroups that are derived from the innovative changes, not the other way around.

[xvii] For instance, in the single sample given by Greenhill & Gray (2012:379) to exemplify coding cognacy, using the word for 'bone', Manggarai *toko* is incorrectly treated as belonging to the same set as Bare'e *wuku*. However, Bare'e *wuku* comes from PMP *bukuh 'node, joint, know' (Blust & Trussel 2020) while Manggarai *toko* comes from an unrelated etymon whose cognates are all local to the Flores area. Despite the input of experts, there remain many such errors in the underlying database.

[xviii] A further oddity of phylogenetic computational models is that they rely on linguists to identify loans and cognates based on the comparative method before the Bayesian analysis can be run. It is thus hard to avoid Donohue et al.'s conclusion (2012b: 544) that the method "uses the results of the comparative method to weakly emulate comparative method results". Similarly, Dyen (1953), using a post-hoc lexicostatistical replication, claimed to "confirm" Dahl's finding that Malagasy subgroups closely with the Bornean language Maanyan although it was unable to arrive at this discovery independently.

[xix] Yoon (2009) and Miller (2020) provide two excellent popular accounts of these developments in biology.

[xx] Note that the same could be said for machine translation, where the statistical black box of "Deep Learning" has garnered far more interest and funding than the more transparent methods devised by linguists. In the field of machine translation, however, the results speak for themselves and Deep Learning algorithms currently outperform all other approaches.