Statistical Reasoning in the Evaluation of Typological Diversity in Island Melanesia

Michael Dunn,*† Robert Foley,‡ Stephen Levinson,†
Ger Reesink,* and Angela Terrill*

*RADBOUD UNIVERSITY NIJMEGEN
†MAX PLANCK INSTITUTE FOR PSYCHOLINGUSITICS, NIJMEGEN
‡LEVERHULME CENTRE FOR HUMAN EVOLUTIONARY STUDIES, CAMBRIDGE

This paper builds on a previous work in which we attempted to retrieve a phylogenetic signal using abstract structural features alone, as opposed to cognate sets, drawn from a sample of Island Melanesian languages both Oceanic (Austronesian) and (non-Austronesian) Papuan (*Science* 2005[309]: 2072–75). Here we clarify a number of misunderstandings of this approach, referring particularly to the critique by Mark Donohue and Simon Musgrave (in this same issue of *Oceanic Linguistics*), in which they fail to appreciate the statistical principles underlying computational phylogenetic methods. We also present new analyses that provide stronger evidence supporting the hypotheses put forward in our original paper: a reanalysis using Bayesian phylogenetic inference demonstrates the robustness of the data and methods, and provides a substantial improvement over the parsimony method used in our earlier paper. We further demonstrate, using the technique of spatial autocorrelation, that neither proximity nor Oceanic contact can be a major determinant of the pattern of structural variation of the Papuan languages, and thus that the phylogenetic relatedness of the Papuan languages remains a serious hypothesis.

1. INTRODUCTION.[1] In an earlier paper (Dunn et al. 2005), we examined the possibility of extracting a phylogenetic signal, and more broadly a signal of historical relatedness, from purely formal (phonological and morphosyntactic) features of languages. Such a method would be independent of historical relations found by the comparative method or other vocabulary methods, and it might apply where for one reason or another (e.g., vocabulary loss) the comparative method cannot. Applying modern computational

1. This work was supported by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek, by means of a Programma grant (Dunn and Reesink) and a Vidi grant (Terrill); support from the Max Planck Institute for Psycholinguistics is also gratefully acknowledged. We would like to thank Åshild Næss for updating her typological questionnaire for this paper on the basis of her Äiwoo fieldwork. We thank Russell Gray, Simon Greenhill, Fiona Jordan, Asifa Majid, and Doug Marmion for comments on earlier versions of this paper. We also appreciate the fact that Mark Donohue and Simon Musgrave made their paper available to us for comment prior to publication.

methods for tree reconstruction to these purely typological features, we found that we could to a very high degree recapitulate the comparative method tree for a branch of the Oceanic languages. We then turned the methods on an analysis of the offshore Papuan languages, too separate in space and time to show any convincing cognates, and were able to recover plausible suggestions for ancient relatedness.

1.1 A RESPONSE TO DONOHUE AND MUSGRAVE. Donohue and Musgrave (2007) find fault with this paper on a number of grounds, querying methods, assumptions, and data. Much of this criticism reflects a misunderstanding of the statistical nature of the methods, and a lack of appreciation of their robustness to small amounts of error in the data. While answering some of the charges in detail, we take the opportunity here to justify our main thesis, namely that the computational reconstruction of language history using typological features offers a new and exciting prospect for understanding language prehistory. To this end we will:

- demonstrate that typological data can be used to show that the non-Austronesian (Papuan) languages in the Dunn 2005 study form a valid historical group, and
- show how phylogenetic methods can be used to make valid claims about linguistic relatedness on the basis of typological evidence.

To this end we will also discuss the advantages of model-based phylogenetic inference for investigating linguistic change using realistic assumptions.

Beyond these general points, this paper also constitutes a concrete response to some of the misapprehensions about Dunn et al. (2005) found in Donohue and Musgrave (2007) (henceforth D&M). The most important claim made by D&M is that the apparent phylogenetic pattern might be a product of areal factors, in particular, contact with Oceanic languages. This is something that we have devoted particular attention to (Dunn et al. 2005:2074, Dunn to appear); in 4.2 we present a spatial autocorrelation analysis that shows that it is very unlikely that Oceanic contact has contributed much to the pattern of typological diversity of the Island Melanesian Papuan languages.

D&M's criticism focuses on assessing the kind of typological data used, and its robustness. In this paper we will discuss further our language sample (2.1) and our choice of typological features (2.2). D&M find fault with a number of the codings of specific features we used in the original paper. While we acknowledge that our data is not perfect, we can demonstrate the robustness of our results by comparing our analysis of the original data with a reanalysis of this data (using Bayesian methodology) with all of the features criticized by D&M removed.

Section 3 overviews phylogenetic modeling of language evolution. In 3.1 we give a justification for a broader understanding of what constitutes the "history of a language", and in 3.2 present a prose description of phylogenetic inference using formal models of evolutionary change. In 3.3 we discuss some differences between the Parsimony method of phylogenetic inference used in Dunn et al. 2005, and Bayesian methods.

Section 4 goes more deeply into the Dunn et al. 2005 data in order to show that investigation of individual typological features is not a profitable way to do linguistic comparison (4.1); establish by means of spatial autocorrelation that "Island Melanesian Papuan" is a valid historical group (4.2); and demonstrate with new data from Äiwoo (Reef

Islands) that aggregate typological comparison can produce useful hypotheses about the affiliation of individual languages (4.3).

1.2 WHAT WE DO NOT CLAIM. We do not claim to have shown that the Island Melanesian Papuan languages have one linguistic ancestor, that is, form an unequivocally defined language family. Nor do we claim that Dunn et al. (2005) is the only or even the best implementation of these techniques (it was an early attempt at the phylogenetic analysis of typological variation). In particular, the use of the parsimony method is not necessarily the best tool for the job. We used it because it is conceptually simple and, in the opinion of some, methodologically conservative. However, Bayesian likelihood methods have been shown to be both more likely to detect a phylogenetic signal where it exists, and less likely to produce a false positive (Ronquist 2004). We apply Bayesian methodology in 2.2 in demonstrating the robustness of the signal in our data to various analytical techniques.

2. DATA

2.1 SAMPLE OF LANGUAGES. The sample of fifteen Papuan languages was dictated by the availability of descriptive materials. Research was carried out under the ESF/OMLL "Pioneers of Island Melanesia" grant, which included fieldworkers who worked on more than half of this sample: Savosavo (Wegener), Lavukaleve (Terrill), Touo (Dunn and Terrill), Rotokas (Robinson), Kuot (Lindström), Kol (Reesink), Sulka (Reesink), and Yélî Dnye (Levinson). Of the remaining languages, grammatical descriptions were available for a few (e.g., Motuna, Bilua); fieldworkers for some others collaborated with us in filling out questionnaires (Mali, Anêm, Ata); and a sufficient number of descriptive papers were available for some others (Buin, Nasioi, etc.). Some materials were available for languages like Nagovisi and Kaket Baining, but we judged them insufficiently complete or insufficiently reliable. Næss has provided us with an Äiwoo questionnaire based on fieldwork she carried out after the publication of Dunn et al. (2005); this will be discussed in 4.3 in the context of the hitherto debated genealogical status of Äiwoo as an Austronesian or non-Austronesian language.

Most of the Papuan languages are not geographically adjacent, but interspersed by descendants of the more recent influx of Austronesian speakers. These Oceanic Austronesian languages provide a useful, but not in all ways ideal, experimental control for the methodology that we are developing. It is useful (although not crucial to the validity of the test) that these languages are distributed over the same general geographic range, and it is necessary for their use as a control that they have a reasonably well established phylogeny that can be compared to our test results. However, the relative uniformity of these languages and the relative youth of the group makes direct comparison difficult.

Obviously, most immediate neighbors of the Island Melanesian Papuan languages belong to the Meso-Melanesian linkage of the Oceanic subfamily, but it was important to have a reasonable representation of the North New Guinea and Papuan Tip linkages as well, because without a genealogically balanced sample of languages we lack a robust *target phylogeny* for comparison. In a more recent and much more detailed comparison

(Dunn et al. MS), an expanded sample of Oceanic languages is compared to the same sample of Papuan languages. In that work, we also trace the position of Mussau and the Admiralties family, and various putative North New Guinea languages of New Britain, because these languages posed problems for the comparative method employed by Lynch, Ross, and Crowley (2002:98).

The sources we consulted for the Oceanic languages were as much as possible from Lynch, Ross, and Crowley (2002). Of course we were aware of the existing full descriptive grammars of Oceanic languages. The short sketches we consulted provided clear, unequivocal accounts with the same theoretical orientation, produced under consistent and high quality editorial oversight. This facilitated the task of coding such a large data matrix by a relatively small team in relatively short time.[2] For what it's worth, there is no statistically significant correlation between the length of a grammatical description in our sample and the completeness of the questionnaire that we filled out from it. In fact, the length of the description predicts less than 2% of the variation in the completeness of questionnaires ($r^2 = 0.019$, $p = 0.6$ for the counts given by D&M). Short sketches are not necessarily less useful than longer descriptions.

While this paper is concerned solely with the (re)analysis of the Island Melanesia data, work is in progress on a wider sample of Papuan, Austronesian, and Australian languages (Reesink and Dunn 2007).

2.2 CHOICE OF FEATURES. The choice of typological features was not just informed by the findings for the East Papuan languages in Dunn, Reesink, and Terrill (2002), but also by the general typological overviews of Papuan versus Austronesian languages in Foley (1998, 2000) and the specific typological characterization of Oceanic languages in Lynch, Ross, and Crowley (2002). The suggestion by D&M that the features selected would somehow be biased by "the subset of the [Dunn et al.] 2005 authors" is unfounded.

Because it is well known that Island Melanesia has been the scene of long-term contact between Oceanic languages and the heterogeneous Papuan languages, we purposely included features that had been identified to cross linguistic lines. As is widely recognized by now, virtually any linguistic trait can be borrowed. Speakers in a bilingual situation can transfer properties of any domain of a grammar from one language to another, whether these languages share a common ancestor or not. Not all kinds of lateral transfer occur with equal ease, but it is impossible to exclude a priori the possibility of any particular change from happening (Thomason 2001).

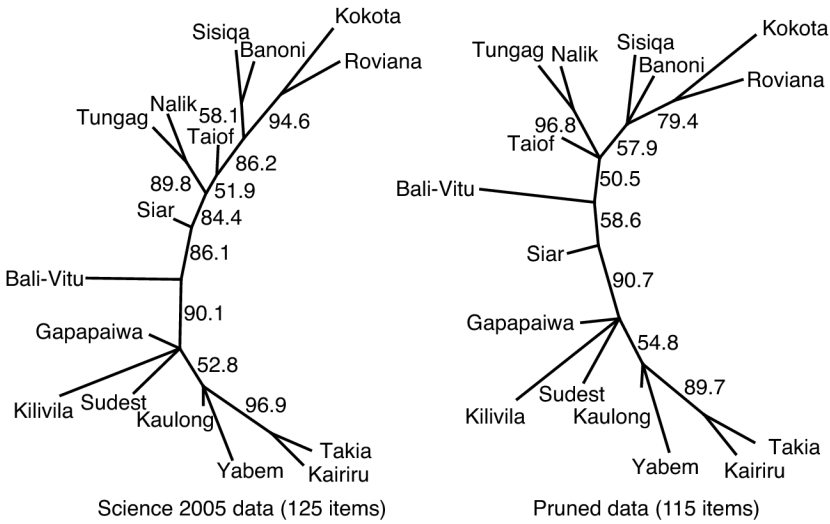
D&M make a number of significant criticisms of particular items in the typological questionnaire itself. Their point is well taken that some of the typological features in the questionnaire remain typologically or logically dependent, despite our attempts (noted in the original paper) to minimize this. Typological dependence can be a problem for the "Parsimony method," the phylogenetic method used in the Dunn et al. (2005) paper, although the extent that the particular typological dependencies bias the results in the

2. D&M invoke the WALS (Haspelmath et al. 2005), but this database is not constructed in such a way as to be easily used in computational phylogenetic analysis: the coding of features is idiosyncratic (e.g., not explicit or exhaustive), many features are inappropriate for phylogenetic analysis, and the data matrix itself is sparsely filled, especially in the region we are interested in.

actual questionnaire used is uncertain, and not likely to be great. One of the attractions of adopting model-based likelihood methods, such as Bayesian phylogenetic inference, is that dependencies between features have less effect on the analysis.

We can use this Bayesian technique to make a simple demonstration of the robustness of structural phylogeny: we can remove all features criticized by D&M, and compare the results of the analysis of this pruned data set with the results of the analysis of the original data. Removing the features criticized by D&M reduces the number of features coded per language from 125 to 115. The results from the two analyses (i.e., plus or minus the contested features) are shown in figure 1: these are the two 50% consensus trees[3] of the Bayesian tree sample from analyses carried out (a) with the original and (b) with the pruned data sets. The differences are minimal. The probability estimates on each branch differ somewhat, but these will always differ slightly even between analyses using the same input data. The only major difference is the reversal of the order of Bali-Vitu and Siar on the tree (disagreeing with the position on the tree arrived at by Lynch, Ross, and Crowley 2002 on traditional, vocabulary-based methods); but note that this new position is inferred with a lower level of certainty than the "correct" (i.e., comparative method) estimation of the position shown on the (a) tree.

FIGURE 1. CONSENSUS TREES PRODUCED BY BAYESIAN PHYLOGENETIC INFERENCE METHODS FROM (A) THE ORIGINAL DATA, AND (B) THE "PRUNED" DATA SET FOLLOWING EXCLUSION OF CHARACTERS CRITICIZED BY D&M



3. A *50% consensus tree* is a tree constructed from a set of trees (in this case, a set of equally probable phylogenetic inferences) by tabulating all the two-way splits that occur in more than 50% of the tree set, and adding them to an unresolved (star-like) consensus tree in order of frequency starting from the most frequent, while skipping those that conflict with a bifurcation already added. A *consensus network* (Holland et al. 2004, Dunn to appear) is a richer representation of the phylogenetic signal in the tree set, but in this test of the sensitivity of the analysis to variation in the data the result is clear without this extra complexity.

A broader observation to make at this point (and one that will be reiterated in more detail below) is that a probabilistic model of evolution with quantified uncertainty is robust against a certain amount of error in the database.

One caveat about the Oceanic test as a validation of the method is that the Oceanic tree itself has considerable uncertainty. The comparative method has not produced a conclusive tree for a number of parts of Oceanic (this has necessitated the introduction of non-standard terminology such as the "linkage", a group linked by overlapping subsets of a set of innovations). Further, the phylogenetic reconstruction of Austronesian using the *cognate-birth, word-death* model, consistently fails to identify a Western Oceanic clade, despite producing a tree otherwise highly congruent with the comparative method (Greenhill and Gray 2005). But note that the uncertainty in the reference form of the Oceanic tree does not effect the demonstration of the robustness of the tree inference presented above in figure 1.

3. PHYLOGENETIC METHODS

3.1 THE SCOPE OF HISTORICAL LINGUISTICS. Traditionally, the scope of historical linguistics includes efforts to identify instances of genetic relatedness between languages, to explore the history of individual languages, and—perhaps of most interest to linguists themselves—to develop a theory of linguistic change. The most important work of lasting value has been done within the paradigm of the comparative method, which models language change as a series of uniquely identifying "mutation" events (for the most part regular sound changes) that define bifurcations in a phylogenetic tree. While language change has always been modeled as an evolutionary system, modern phylogenetic methods allow a wider range of different aspects of language to be modeled. Statistical, computational, and algorithmic work on evolutionary trees is barely forty years old, and the application of these methods outside biology (e.g., in language and culture) is only in its early stages. Adoption of these new methods has enormous scientific advantages, allowing modeling of more realistic assumptions about the processes of language change, and producing results with a number of statistically interesting and highly informative properties such as quantified uncertainty and relative chronology.

The comparative method in historical linguistics presumes a strict notion of what counts as a historical signal. In an important recent paper in the *Handbook of historical linguistics*, Harrison makes explicit that the business of historical linguistics is to investigate vertical transmission, what Harrison refers to as "normal historical continuation": "Being more precise about what is meant by "normal historical continuation" isn't easy. It must involve notions like "normal language acquisition" and "normal language change." Although there may be some danger of circularity here, it seems to me safe to assume that historical linguists will know what I have in mind" (Harrison 2003:217).

But the consensus on what counts as normal language change is broadening. The idea of what is "normal" referred to by Harrison above could be seen as a pragmatically motivated compromise, such that "normal" linguistic change is at least in part determined by the kinds of linguistic change that are tractable using the comparative method. However, it is hardly controversial to acknowledge that language contact, horizontal processes of

language change such as substrate effects, code-switching, and so on are the norm throughout most language communities, present and past. If we accept a more liberal notion of what is normal in linguistic change we are forced to question the validity of the strict partition between vertical and horizontal processes of change, between inheritance and contact. Methods that include more realistic models of language change can form the basis of more interesting inferences about linguistic prehistory.

Vertical and horizontal modes of transition form the extremes of a cline, and different linguistic subsystems behave differently along this cline. Regular sound change is normally associated with vertical transmission, but counterexamples are not infrequent (e.g., the cascade of sound changes in Western European languages following from the horizontal spread of the uvular *r* from the Romance subgroup to Germanic); some typological features are considered to be extremely susceptible to horizontal transfer, yet in some cases plausible arguments for an ancient phylogenetic signal can be made (Dunn et al. 2005). Lexeme history, as modeled by lexical innovation and loss (the *cognate-birth, word-death* model used by e.g., Gray and Atkinson 2003) inhabits an intermediate position. These various linguistic subsystems allow reconstruction of different kinds of sociolinguistic processes (e.g., where particular processes can be shown to favor vertical or horizontal transmission types). Hypotheses of "abnormal" transmission (substrate, language shift, networks) are often invoked to account for unexpected and/or uninterpretable comparative method results. Where computational evolutionary methods can show a mismatch between different linguistic subsystems, reconstruction of complex language and population processes becomes possible.

With the establishment of a statistically sophisticated, probabilistic evolutionary framework we gain the ability to compare data sets from outside of linguistics as well. Spatial autocorrelation and other geostatistical techniques allow testing of dispersal hypotheses (for example Hunley et al. 2007). Analogously to the different histories of linguistic subsystems, mtDNA and Y-Chromosome DNA histories track different genetic processes in human populations (the matrilineal and patrilineal histories respectively), and these also may correlate with linguistic variation. For example, linguistic typological variation in New Britain (Papua New Guinea) correlates best with mtDNA variation and not at all with Y-Chromosome DNA, presumably reflecting greater male migration across language groups than female (Hunley et al. 2007).

3.2 MODELING THE EVOLUTION OF LANGUAGE. Inferring the history of linguistic change is always an exercise in modeling the evolution of language. The comparative method has an implicit model of language evolution in which change is characterized as unique phonological "mutations" (shared regular sound changes). The comparative method is non-statistical, and so does not consider, for example, the number of words showing evidence of these sound changes. On the more ancient nodes of a comparative method tree there may be only two or three words showing evidence for a sound change, the rest of the lexicon having been subsequently replaced. The current statistical models of language evolution treat languages as a collection of traits; the model is actually a model of the behavior of these traits: binary (0-1, yes-no, present-absent) traits may be characterized as the probability of a state change per unit time (this is *relative time*,

as measured by distance on the tree, not calendar time). There are many possible elaborations on such a model: for example, state change probability may be allowed to differ for gains (0 1) and losses (1 0), and traits may be allowed to belong to different rate change categories (each characterized by different gain and loss probabilities).

Likelihood models (including Bayesian) are currently the most promising methods for phylogenetic inference (Holder and Lewis 2003). The methods implement a tree/parameter search, where the best phylogeny is the one that is most likely to produce data like the observed data. This was implemented in 2.2 (figure 1).

Algorithm and software development in computational phylogenetics has produced a number of implementations of analytic tools using Bayesian phylogenetic inference (we use BayesPhylogenies, Pagel and Meade 2004). Bayesian methods in general are valuable whenever there is a need to extract information from data that are uncertain or subject to any kind of error or noise (including measurement error and experimental error, as well as noise or random variation intrinsic to the process of interest). Bayesian phylogenetic inference allows quantification of uncertainty and the investigation of conflicting phylogenetic signals (see Huelsenbeck et al. 2001).

In Bayesian phylogenetic inference, taxa (in our case, languages) are modeled as a set of characteristics. These characteristics (*characters* in phylogenetic terminology) have their own probabilities of change. A hypothetical phylogeny for a group of taxa includes a genealogical tree—a hypothesis about tree topology, including branch lengths (showing relative chronology)—and also the hypothetical values for the set of transition probabilities (the probabilities of the characters changing back and forth into their different states). A set of transition probabilities can in principle produce *any* tree, but some trees are much more likely than others. The task of Bayesian phylogenetic inference is one of optimizing the parameters of the evolutionary model, that is, to discover for a set of observed states the most likely parameters to have produced them. The details of the optimization procedure used to search the set of possible parameters for the most likely ones is too complex to set out here (but see Dunn [to appear] for a nonmathematical account), but note that the parameter search encompasses the entire model, which means we are simultaneously searching for *both* the transition probabilities of the characters *and* the tree that they produce. The result of this parameter space search is—where the data contains a detectable phylogenetic signal—a sample of equally likely trees.

In our work on the Papuan languages of Island Melanesia, we have focused on modeling the evolutionary processes of typological, rather than lexical, traits. There is no a priori reason to think that this would work, but likewise there is no a priori reason to assume that typological traits would carry no phylogenetic information. Dunn et al. (2005) presented the first evidence that there can be phylogenetic information in linguistic typological data. Later work on the Island Melanesian data has provided further confirmation of this. Part of our program for the future is to investigate under what circumstances typological information about languages will tend to preserve phylogenetic information.

3.3 PARSIMONY METHOD. The parsimony method, as used in Dunn et al. (2005), is also a statistical method. In it, an algorithm is used to calculate the phylogenetic tree that requires the fewest number of state changes (i.e., changes in the value of a feature

of a language). It has a number of weaknesses that make it inferior to Bayesian phylogenetic inference (see also Pagel 1999). Most serious perhaps is the phenomenon of *long branch attraction* (Felsenstein 1978). Long branches (representing in our case languages that have had the most amount of change from the reconstructed ancestor node) tend to merge. On long branches a character state may have flip-flopped several times, but the most parsimonious account is to (incorrectly) postulate a shared ancestor to the long branches, so that each taxon needs at most one state change from the ancestral state. Other disadvantages of the parsimony method are that it is not strictly probabilistic, and that it does not produce confidence values (the robustness of a parsimony analysis is usually evaluated using a *bootstrap test*,[4] but bootstrap scores are not probabilities, and cannot be compared between different sets of data).

D&M are incorrect in their description of the treatment of missing data in a parsimony analysis: taxa are never grouped together on the basis of missing data in common. In PAUP* (Swofford 2003), missing data is given the most parsimonious character state for its position in the tree, meaning that only characters with no missing data will affect the placement of taxa (Wiens 2003:531, Swofford n.d.). Wiens (2006:41) concludes a review of the treatment of missing data in phylogenetic analysis (including, among others, parsimony and Bayesian phylogenetic inference): "Recent simulations show that there is little evidence to support excluding taxa based simply on the amount or proportion of missing data that they bear. The placement of highly incomplete taxa in a phylogeny can be resolved with perfect accuracy (based on simulations) and with strong statistical support (based on empirical analyses). The critical factor determining their placement is seemingly the characters that are present in these taxa (i.e., their number and quality) not the ones that are absent."

4. AUSTRONESIAN AND NON-AUSTRONESIAN LANGUAGES

4.1 THE FOLLY OF ESSENTIALISM. D&M make a great deal out of the weakness of the proposed method for distinguishing two language families within a combined sample. Their argument echoes a common misapprehension that individual features within our framework can be treated as diagnostic of membership of the Oceanic or Papuan populations.

The division of features into *significant*[5] and *insignificant* (D&M table 6 and §5.2) betrays a serious misunderstanding of the statistical nature of computational phylogenetics. It should not be supposed that the distribution of individual features is a sufficient diagnostic of language families. As established by Thomason and Kaufman (1998), just about anything can be borrowed under the right, often very rare and special, circumstances. Computational phylogenetic methods consider the joint behavior of all features

4. A bootstrap test is a statistical test of the robustness of data in which the analysis is repeated many times using the values from a randomly selected subset of features. If many features contribute to reconstructing a particular branch in the tree, this branch will be present in many of the bootstrap repetitions. If only a few features contribute to the branch, then it is likely to be absent from most of the bootstrap set. The bootstrap score on a branch is the percentage of the bootstrap set that has that branch. The bootstrap score is *not* a measure of time depth, as implied by D&M.

5. D&M use the term *significant* in both the statistical and the non-technical sense. *Caveat lector!*

(cf. Nichols 1996), and the probability of independent acquisition of shared features drops exponentially with the number of features. We do not presume to address the remarkable array of typological facts that D&M have marshaled in their paper, except to say that it is precisely because there are exceptions to the general associations of particular typological features with particular linguistic groups that a statistical, probabilistic method is required. We certainly do *not* expect any single feature to be diagnostic of a family, and we do not expect the search for diagnostic features by inspection to be profitable. Identification of patterns in complex data sets is something that humans do with great, but frequently misplaced, confidence (Kahneman, Slovic, and Tversky 1982).

Counterexamples, where D&M's supposed diagnostic features occur in the *wrong* language group are likewise of little value. In anthropology this is called "bongo-bongoism" (Douglas 1970:15-16), defined as "the trap of all anthropological discussion. Hitherto when a generalization is advanced, it is rejected out of court by any fieldworker, who can say 'this is all very well, but it doesn't apply to the Bongo-Bongo'." Linguistic typologists are no less prone to bongo-bongoism than anthropologists: we treasure our every hard-earned field observation, and rightly respect our peers for their broad knowledge of linguistic diversity. However, a few features out of place does not invalidate a general trend: if what we do is to be a science, there comes a point in the process of an analysis when we have to treat linguistic data statistically, as an aggregate. The list of examples and counterexamples produced by D&M has value as the beginnings of a richer database, but does not in its present form offer us anything useful for evaluating the Dunn et al. 2005 study.

4.2 ESTABLISHING THAT PAPUAN IS A VALID GROUP. It is quite possible that the phylogenetic analysis of the typological features of a handful of Oceanic and Island Melanesian Papuan languages might not clearly establish the membership and genealogy of the two groups.

For there is no reason to suppose that a phylogenetic analysis of language data from two different families would come up with a sensible tree. Dumping all the languages you know about into a phylogenetic analysis is the wrong strategy to look for divisions into populations (there are other computational techniques that are better suited to that task, e.g., the program STRUCTURE; cf. Pritchard, Stephens, and Donnelly (2000), Rosenberg et al. 2002). A phylogenetic inference comes up with the most plausible evolutionary pathway to produce the observed linguistic variation, assuming that all the taxa under consideration belong to a single genealogy. The problem is that if there are two families in the data, the protolanguages of the two families are quite likely to be more dissimilar than some other unrelated pairs of languages in the sample; and thus forcing the two families into a single phylogeny will tend to join the two groups far from their true root nodes. The way out of this procrustean bed is to divide the data into two groups prior to analysis (either on the grounds of ancillary evidence, or using other computational techniques like NeighborNet, cluster analysis or STRUCTURE; see Dunn et al MS), and to perform independent tests of the coherence and phylogenetic plausibility of these groups.

Oceanic is solidly established by lexical comparison; but *given* the existence of an Oceanic group, we can seek to demonstrate the plausibility or otherwise of an Island Melanesian Papuan phylogenetic group. The evidence given in Dunn et al. (2005) hinged on the

seemingly greater than chance congruence of the phylogenetic tree to geographic distribution. However the geographic distribution of typological variation can be investigated more rigorously than this, and the results of such investigation strongly support our original hypothesis, that the Papuan languages show some kind of ancient historical relatedness.

The analysis (known variously as *spatial autocorrelation* or *isolation by distance*) compares the geographic separation of pairs of languages to their degree of structural difference. We will refer to this as "geographic" and "structural" distance. Geographic distance is measured in kilometers via waypoints set between the major island groups. In the extreme case this changes the circa 600km between Lavukaleve and Yélî Dnye "as the crow flies" to more than 2000km "as the canoe sails." (The use of waypoints slightly improves the predictiveness of the model, but the results are strong even when using direct distances.) We plot these distances on a logarithmic scale, so that the effect of very great distances eventually flattens out. Structural distance is measured as the proportion of valid pairs of feature values that differ between the pair of languages (i.e., the number of values that are different out of all the features where both languages have a *yes* or a *no*; features where either language of the pair has *unknown* are excluded from the distance measurement). In normal cases of evolutionary diversification, taxa tend to be phylogenetically more closely related to those taxa that are also geographically proximate. If this is the case then a scatterplot plotting geographic distance against structural distance will show a generally upward trend. We can model this by a line. The slope of the line (the rate at which structural distance increases with geographic distance) is interesting, but not as important as the squared correlation coefficient (the r^2 statistic), which tells us what proportion of the variance of the scattered points is accounted for by the line (so that if all points were perfectly arrayed along the line the r^2 value would be 1.0, whereas the r^2 of widely scattered points will approach 0).

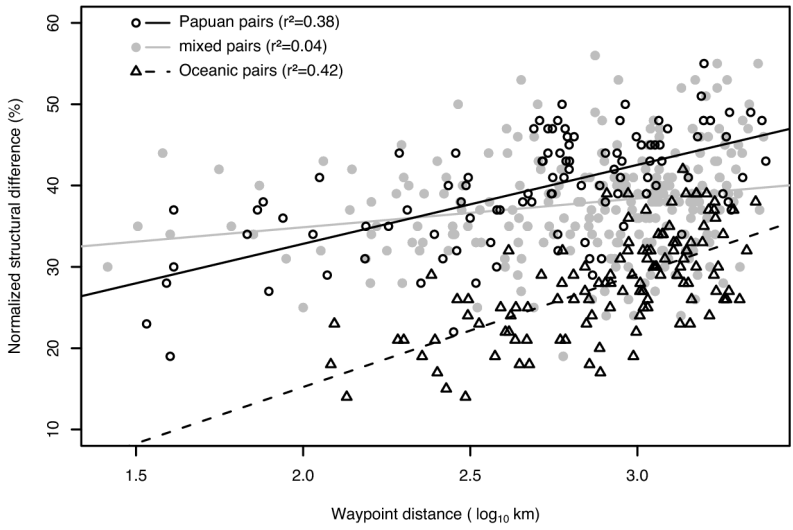
Figure 2 shows comparisons of pairs of Oceanic languages, pairs of Papuan languages, and "mixed" pairs, that is pairs containing one Oceanic and one Papuan language. The Oceanic–Oceanic and Papuan–Papuan pairs both show a strong signal of isolation by distance (with the linear model accounting for about 40% of the variance in each case). That is, the geographically closer a pair of languages is, the more structural features they share. This goes only for Oceanic–Oceanic pairs, or Papuan–Papuan pairs. If the structural similarities between the Papuan languages were the result of Oceanic contact we would expect that the mixed Papuan–Oceanic pairs would show isolation by distance as well. That is, the geographically close Papuan–Oceanic pairs should share more structural features than geographically further apart Papuan–Oceanic pairs. But this is not the case: geographic proximity accounts for only 4% of the typological diversity of mixed Papuan–Oceanic pairs. A Papuan language is hardly more likely to share structural similarities with a nearby Oceanic language than a distant one. This shows that the signal of similarity between the Papuan languages cannot be accounted for by contact-induced language change from Oceanic.

4.3 THE GENEALOGICAL POSITION OF ÄIWOO. There has been considerable debate about the status of Äiwoo as an Austronesian or non-Austronesian language (the former position originally argued for by Lincoln, the latter vigorously

defended by Wurm). We did not consider Äiwoo in the 2005 study, as there was not adequate published information to fill out a questionnaire, and in any case, we did not consider that the affiliation of Äiwoo was reliably established (contra D&M, who criticize the 2005 study for omitting Äiwoo from the Papuan sample). Since the original study was carried out, Næss has begun to publish Äiwoo data on the basis of new fieldwork (Næss 2006), and has completed a typological questionnaire (Åshild Næss, pers. comm.). Reesink (2006) presented the NeighborNet[6] clustering of Äiwoo (figure 3), showing the position of Äiwoo among a sample of the languages of Bougainville and Solomon Islands. Äiwoo patterns strongly with Buma (Vanikoro Island), the nearest (non-Polynesian) Oceanic language in the sample. Note also the very strong division between Austronesian and non-Austronesian languages in this network.[7]

The strong patterning of Äiwoo with Austronesian languages, rather than the other non-Austronesian languages of Bougainville–Solomons, is consistent with Næss (2006), a paper that clarifies the essentially Austronesian behavior of the purported noun classes, as well as Ross and Næss (2007), which gives the definitive comparative method classification of Äiwoo as a member of a first-order subgroup of Oceanic. The NeighborNet

FIGURE 2. SCATTERPLOT SHOWING THE CLUSTERING OF GEOGRAPHIC VS. STRUCTURAL FEATURES OF ALL PAIRS OF LANGUAGES IN THE DATA



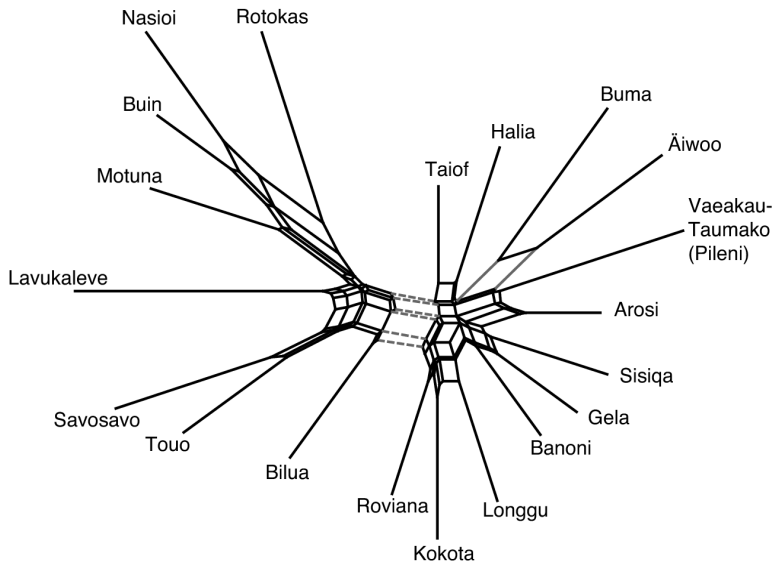
6. NeighborNet (Bryant and Moulton 2003, Bryant, Filimon, and Gray 2005) is a method for displaying binary divisions in data while not concealing conflicting evidence. It is not a *phylogenetic* method: it converts aggregate data to distances, and so does not model evolutionary processes. Nevertheless, it is a useful first step in an analysis, because the more treelike a NeighborNet graph is, the more phylogenetic signal the data are likely to contain.
7. The solid gray lines show a split between the two Temotu languages (excluding Vaeakau-Taumako, a Polynesian back-migration to Temotu) and the rest; dashed gray lines indicate the strong split between Austronesian and non-Austronesian.

graph in figure 3 clearly shows that the typological profile of Äiwoo is in no way aberrant for an Austronesian language located in Temotu.

5. CONCLUSION. We take it as axiomatic that languages embody more than one historical source. In the same way that modern genetics traces different aspects of human history by the Y-chromosome DNA, mitochondrial DNA, or autosomal recombinant DNA, the evolutionary science of language change can address language history by looking at all the different levels of linguistic patterning (sound changes, vocabulary sets, typological features, and presumably other phenomena too). Furthermore, the distinction between vertical and lateral transfer is not as sharply delineated as some traditional models of language change presume, and lateral transmission should be included in a theory of language change as much as vertical transmission.

Where traditional methods of analysis, concentrating on lexical comparisons, fail to yield any historical picture, it would be a mistake to look no further. Island Melanesia is a good test ground for such new methods. Although lexical and phonological methods cannot detect relationships between the Papuan languages of Island Melanesia, it is plausible from the point of view of history and geography that some kind of historical relationships nevertheless exist between them, because these languages are remnants of

FIGURE 3. NEIGHBORNET GRAPH OF BOUGAINVILLE AND SOLOMON ISLAND LANGUAGES



linguistic diversity that existed before the relatively recent incursion of the Austronesian languages (otherwise we must suppose 26 far-flung independent migration events for each of the Papuan languages). The strong correlation between the spatial distribution and the structural diversity of the Papuan languages discussed in 4.2 shows unquestionably that some kind of historical relationship between these languages does indeed exist. The lack of correlation between spatial and structural distance for mixed pairs of Papuan and Austronesian languages shows that the relationships between Papuan languages cannot be merely the reflection of shared patterns of Austronesian contact. This leaves us with three possibilities: the relationships between the Papuan languages may be the product of ancient intra-Papuan contact, they may be the result of common ancestry, or it may be the case that both phylogeny and ancient contact have contributed to the relationships between the Papuan languages of Island Melanesia.

We are unable to tease ancient contact and phylogeny apart, but we can at least begin the process. One question we can ask is: assuming that the relationship between some Papuan languages is phylogenetic in nature, what would this phylogeny look like? Our experiments with recapitulating Austronesian comparative method phylogenies using typological data and Bayesian phylogenetic inference suggest that the Papuan phylogeny based on structural features presented in Dunn et al. (MS) is a plausible hypothesis that should guide further research in this area. This research program is young, and there is certainly much more to be done.

As a final methodological observation, we would also like to point out that the methods of structural phylogeny are useful not only for genealogical reconstruction. Rigorous model-based evolutionary analysis of language change allows us to study the behavior of individual typological traits within a family, so that we will be able to make statements about the stability of structural features (see Gray and Atkinson 2003, Pagel and Meade 2006 for these techniques used with lexical characters), and investigate correlated evolution between pairs of features (Dunn and Greenhill 2007).

REFERENCES

Bryant, David, and Vincent Moulton. 2003. Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution* 21:255–65.

Bryant, David, F. Filimon, and Russell Gray. 2005. Untangling our past: Languages, trees, splits and networks. In *The evolution of cultural diversity: Phylogenetic approache,*. ed. by Ruth Mace, Clare J. Holden, and Stephen Shennan, 67–84. London: UCL Press.

Donohue, Mark, and Simon Musgrave. 2007. Typology and the linguistic macro-history of Island Melanesia. *Oceanic Linguistics* 46:

Douglas, Mary 1970. *Natural symbols*. New York: Pantheon Books.

Dunn, Michael. To appear. Contact and phylogeny in Island Melanesia. *Lingua*.

Dunn, Michael, Ger Reesink, and Angela Terrill. 2002. The East Papuan languages: A preliminary typological appraisal. *Oceanic Linguistics* 41:28–62.

Dunn, Michael, Angela Terrill, Ger Reesink, Robert A. Foley, and Stephen C. Levinson. 2005. Structural phylogenetics and the reconstruction of ancient language history. *Science* 309:2072–75.

Dunn, Michael, and Simon Greenhill. 2007. Correlated evolution of structural features of languages. Paper presented 17 June 2007 at COOL7, Noumea.

Dunn, Michael, Stephen C. Levinson, Eva Lindström, Ger Reesink, and Angela Terrill. MS. Structural phylogeny in historical linguistics: Methodological explorations applied in Island Melanesia.

Felsenstein, Joseph. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology.* 27:401–10.

Foley William A. 1998. Toward understanding Papuan languages. In *Perspectives on the Bird's Head of Irian Jaya, Indonesia*, ed. by Jelle Miedema, Cecilia Odé, and Rien Dam, 503–18. Amsterdam: Rodopi.

———. 2000. The languages of New Guinea. *Annual Review of Anthropology* 29:357–404.

Gray, Russell D., and Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426:435–9.

Greenhill, Simon, and Russell D. Gray. 2005. Testing population dispersal hypotheses: Pacific settlement, phylogenetic trees and Austronesian languages. In *The evolution of cultural diversity: A phylogenetic approach*, ed. by Ruth Mace, Clare Holden, and Stephen Shennan, 31–52. London: UCL Press.

Harrison, Sheldon P. 2003. On the limits of the comparative method. In *The handbook of historical linguistics*, ed. by Brian D. Joseph and Richard D. Janda, 213–43. Oxford: Blackwell.

Haspelmath, Martin, Matthew S. Dryer, David Gil, and Bernard Comrie, eds. 2005. *World atlas of language structures*. Oxford: Oxford University Press.

Holder, Mark, and Paul O. Lewis. 2003. Phylogeny estimation: Traditional and Bayesian approaches. *Nature Reviews Genetics* 4:275–84.

Holland, Barbara, Katharina T. Huber, Vincent Moulton, and Peter J. Lockhard. 2004. Using consensus networks to visualize contradictory evidence for species phylogeny. *Molecular Biology and Evolution* 21:1459–61.

Hunley, Keith, Michael Dunn, Eva Lindström, Ger Reesink, Angela Terrill, Heather Norton, Laura Scheinfeldt, Françoise Friedlaender, D. Andrew Merriwether, George Koki, and Jonathan Friedlaender. 2007. Inferring prehistory from genetic, linguistic, and geographic variation. In *Genes, language, and culture history in the southwest Pacific,* ed. by Jonathan Friedlaender, 141–54. Oxford: Oxford University Press.

Huelsenbeck, John P., Fredrik Ronquist, Rasmus Nielsen, and Jonathan P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science,* 294:2310–4.

Kahneman, Daniel, Paul Slovic, and Amos Tversky, 1982. *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

Lynch, John, Malcolm Ross and Terry Crowley. 2002. *The Oceanic Languages.* Richmond: Curzon Press.

Næss, Åshild. 2006. Bound nominal elements in Äiwoo (Reefs): A reappraisal of the "multiple noun class system." *Oceanic Linguistics* 45:269–95.

Nichols, Johanna. 1996. The comparative method as heuristic. In *The comparative method reviewed: Regularity and irregularity in language change,* ed. by Mark Durie and Malcolm Ross, 39–71. New York, Oxford: Oxford University Press.

Pagel, Mark. 1999. Inferring the historical patterns of biological evolution. *Nature* 401:877–84.

Pagel, Mark, and Andrew Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Systematic Biology* 53:571–81.

———. 2006. Estimating rates of lexical replacement on phylogenetic trees of languages. In *Phylogenetic methods and the prehistory of languages,* ed. by Peter Forster and Colin Renfrew, 173–82. Cambridge: McDonald Institute for Archaeological Research.

Pritchard, Jonathan K., Matthew Stephens, and Peter Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945–59.

Reesink, Ger. 2006. Comments on contact between Papuan and Oceanic. Paper presented at: 'Mapping the Pacific: Positions, Representations and Interpretations' Oceania Seminar, Lysebu, Oslo, 29 September–1 October.

Reesink, Ger, and Michael Dunn. 2007. The quest for historical relations between Papuan languages. Paper given at International Conference on Oceanic Linguistics (COOL7), Nouméa, 3 July.

Ronquist, Fredrik. 2004. Bayesian inference of character evolution. *TRENDS in Ecology and Evolution* 19, 475–81.

Rosenberg, Noah A., Jonathan K. Pritchard, James L. Weber, Howard M. Cann, Kenneth K. Kidd, Lev. A. Zhivotovsky, and Marcus W. Feldman. 2002. Genetic structure of human populations. *Science* 298:2381–5.

Ross, Malcolm and Åshild Næss. 2007. An Oceanic origin for Äiwoo, a language of the Reef Islands. *Oceanic Linguistics* 46:

Swofford, David L. 2003. *PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods), version 4.0b10.* Sunderland, Massachusetts: Sinauer Associates.

———. N.d. *How does PAUP* deal with missing characters under the parsimony criterion?* http://paup.csit.fsu.edu/paupfaq/faq.html. Retrieved 9th August 2007.

Thomason, Sarah G. 2001. *Language contact: An introduction.* Washington: Washington University Press.

Thomason, Sarah G., and Terrence Kaufman. 1988. *Language contact, creolization, and genetic linguistics.* Berkeley: University of California Press.

Wiens, John. 2003. Missing data, incomplete taxa, and phylogenetic accuracy. *Systematic Biology* 52:528–38.

———. 2006. Methodological review: Missing data and the design of phylogenetic analyses. *Journal of Biomedical Informatics* 39:34–42.

Michael.Dunn@mpi.nl

r.foley@human-evol.cam.ac.uk

levinson@mpi.nl

ger.reesink@hccnet.nl

a.Terrill@let.ru.nl