Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Lingua 119 (2009) 1664–1678
Lingua
www.elsevier.com/locate/lingua

Contact and phylogeny in Island Melanesia

Michael Dunn [a,b]

[a] *Radboud University Nijmegen, P.O. Box 310, NL-6500 AH Nijmegen, The Netherlands*
[b] *Max Planck Institute for Psycholinguistics, P.O. Box 310, NL-6500 AH Nijmegen, The Netherlands*

Received 10 April 2007; received in revised form 15 October 2007; accepted 15 October 2007
Available online 9 December 2008

Abstract

This paper shows that despite evidence of structural convergence between some of the Austronesian and non-Austronesian (Papuan) languages of Island Melanesia, statistical methods can detect two independent genealogical signals derived from linguistic structural features. Earlier work by the author and others has presented a maximum parsimony analysis which gave evidence for a genealogical connection between the non-Austronesian languages of island Melanesia. Using the same data set, this paper demonstrates for the non-statistician the application of more sophisticated statistical techniques—including Bayesian methods of phylogenetic inference, and shows that the evidence for common ancestry is if anything stronger than originally supposed.

Keywords: Papuan; Austronesian; Language contact; Linguistic prehistory; Phylogenetic analysis; Bayesian inference

1. Introduction

Island Melanesia is a fascinating natural laboratory for the study of language change. From 30 000 to 3000 years before present the islands of inner Melanesia (the Bismarck Archipelago, Bougainville, and the Central Solomons) formed the furthermost limit of human dispersal into the south Pacific. In the last 3000 years the Austronesian expansion encompassed near Island Melanesia, and continued eastwards into the Pacific (Spriggs, 1997). Currently 90% of the languages of inner Island Melanesia are members of the Oceanic subgroup of the Austronesian family. With some interesting exceptions, these languages fall into reconstructible genealogical relationships using the standard linguistic comparative method. The remaining 10% of the languages of the region are the – hitherto largely unrelatable – Papuan remnants of the pre-Austronesian linguistic diversity.

The Austronesian language family is for the most part a well-defined family according to the comparative method, comprising a nested series of subgroups defined by shared innovations. A large number of phonological, morphological and lexical innovations have been identified defining the Oceanic subgroup of Austronesian (surveyed in Lynch et al., 2002). The homeland of the Oceanic subgroup is thought to be on the north coast of New Britain, and the later dispersal of Oceanic is a multi-directional radiation, unlike the mostly unilinear spread if its Austronesian ancestors. Perhaps because of this the internal subgrouping of Oceanic is less clear, and it may be that a strictly hierarchical tree of relationships within Oceanic will never be possible due to the reticulate nature (i.e., an interconnected network rather than hierarchical tree) of early Oceanic social structure (Lynch et al., 2002:93).

E-mail address: Michael.Dunn@mpi.nl.

0024-3841/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.lingua.2007.10.026

The area in which the Island Melanesian Papuan languages are spoken falls within the geographic range of the Western Oceanic group of Oceanic (Fig. 1). Western Oceanic is a *linkage* (Ross, 1988), a group of languages reconstructed to a dialect network rather than to a single language, with no uniquely defining innovation shared by all of its members. The three subgroups of Western Oceanic treated in this paper are also more properly linkages (Lynch et al., 2002). The dispersal of Oceanic languages through Island Melanesia is thought to have progressed in two major waves (Ross, 1988; Kirch, 1997). The first wave, ancestral to the present day Central/Eastern Oceanic languages, encountered an environment where the major islands were already populated by non-Austronesian peoples, and moved fairly fast, probably via settlements on tiny offshore islets, to the uninhabited islands of outer Melanesia and Polynesia (the current range of CEO languages starts from just to the east of Kokota and extends right across the Pacific to Easter Island). The second wave, ancestral to the Western Oceanic group, spread more slowly, and presumably replaced many languages which were previously present. This might have involved population replacement, assimilation, or language shift; most likely it involved all three in different situations. The human phenotypes in the populations of Island Melanesia are very mixed, and the genetic distinction between Austronesian-speaking versus Papuan-speaking populations is hard to establish (Hunley et al., 2007).

While the Oceanic languages clearly form a genealogical group, the relations between Papuan languages are more problematic. These languages show recurrent typological similarities which suggest relatedness (see, for example, Wurm, 1975; Dunn et al., 2002), but no lexical evidence has been found such that the comparative method can prove relatedness. No shared phonological innovations have been established from the cognate candidates identified by Greenberg (1971); a more recent study by Ross (2001) working with whole pronoun paradigms demonstrates a number of small groups within the Papuan languages, but neither establishes the unity of Island Melanesian Papuan nor attempts a comparative method reconstruction. A controversial proposal suggests that the history of Island Melanesia is so reticulate that discussion of Austronesian and non-Austronesian/Papuans is misplaced (Terrell et al., 2001). While this may be apparent in material culture, and may well turn out to be the case in human biology, the basic linguistic division between Austronesian and non-Austronesian is with rare exceptions unassailable (Bellwood et al., 1995; Adelaar and Himmelmann, 2005). From a quantitative (rather than comparative method) perspective, Greenhill and Gray (2005) have investigated the dispersal pattern of the Austronesian languages through the distribution of cognate retention, and have found a clear phylogenetic signal, congruent with the comparative method tree.

Beyond the areal focus of this paper there is a larger methodological goal. Dunn et al. (2005) presented the first application of computational phylogenetic methods to search for evidence of ancient relationships between the languages of Island Melanesia. This paper uses exactly the same database (see Appendix A), but reanalyzes the data using more sophisticated methods, including Bayesian Phylogenetic Inference and Spatial Autocorrelation. These improvements in the method have corresponded to an increase in confidence in the validity of the original results.

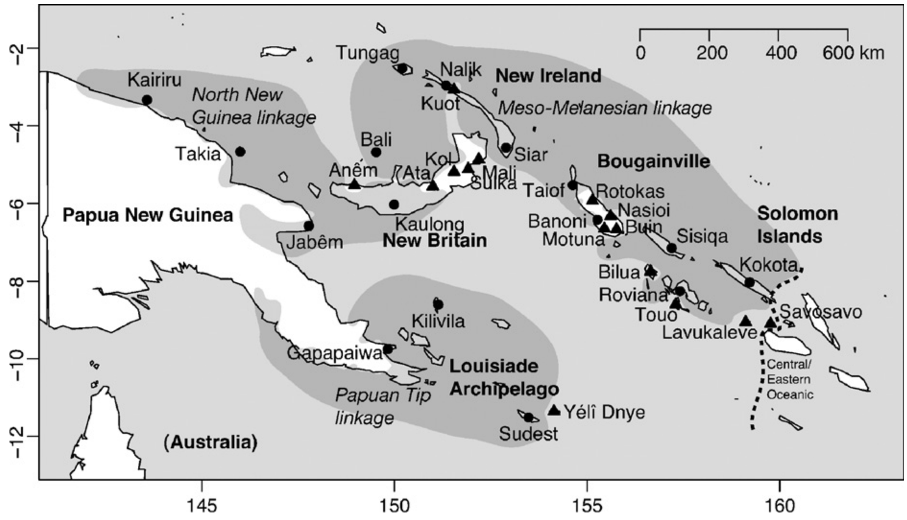

Fig. 1. Map of Island Melanesia showing the approximate locations of the Oceanic languages (circles) and Papuan languages (triangles) used in this study. The three main linkages making up Western Oceanic are indicated in dark gray.

Section 2 describes Bayesian Phylogenetic Inference, presenting an analysis of the Oceanic language sample and a statistical test of the congruence of the resultant tree with the comparative method. Section 3 gives the result of the same phylogenetic analysis with the Papuan language sample, as well as a geostatistical appraisal of typological diversity within and between Papuan and Oceanic groups.

This paper also presents a prose (i.e., non-mathematical) description of the techniques used in the analysis. While this is perhaps anathema to a person equipped to assess the mathematics, it is both possible and useful for the non-mathematician historical linguist to have a high level understanding of what these techniques do and how they do it. Computational phylogenetic methods have their own weaknesses and limitations, and there is no expectation that they will overthrow conventional historical linguistics (McMahon and McMahon, 2005:26–29). But statistical and phylogenetic methods do provide valuable additions to our understanding of linguistic prehistory, including (but not limited to): quantified (un)certainty, branch lengths and relative dating, statistical measures of stability, explicit and realistic models of language change, and integration across disciplinary boundaries.

2. Phylogenetic analysis

In this section, I will work through a phylogenetic analysis of the Oceanic languages in Dunn et al. (2005) database using improved methods, and will discuss how this phylogenetic signal can be evaluated against the reference tree provided by the comparative method. The phylogenetic analysis of the Papuan languages presented in section 3 has been carried out using the same methods.

Due to the continual process of erosion of inherited lexicon in any given language, the phylogenetic signal detectable by the comparative method inevitably disappears into noise at some point. Dunn et al. (2005) present results of a project investigating the linguistic prehistory of Island Melanesian Papuan using computational methods treating structural features (i.e., abstract grammatical features without reference to formal expression) as genealogically transmitted traits. An investigation on the group of Western Oceanic languages gave a control group: a group of languages with a reasonably well known phylogeny from the standard comparative linguistic method. Structural phylogenetic methods use a set of data independent from the formal (largely lexical, with some morphological forms) data of the comparative method, and so the comparative method tree can function as a control to validate the method for inferring the structural phylogeny tree.

There is no doubt that a process of reticulation through borrowing and contact must also explain some of the variation. Any set of linguistic structural data will show some reticulation (i.e., linguistic homology – formal or structural similarities – due to contact-induced change and chance convergence). The NeighborNet method, implemented in the SplitsTree package (Bryant and Moulton, 2004), is useful for visualizing conflicting phylogenetic signals (i.e., signals that show reticulation, violating the expectation of treelike hierarchy; first use in linguistics in Bryant et al., 2005). The NeighborNet network presented in the Supplementary Materials to Dunn et al. (2005) reinforces the observations of descriptive linguists that Oceanic–Papuan contact has had a significant influence on some groups of languages within Island Melanesia. Deeper exploration of the typological database used shows – not unexpectedly – that not all traits are equally stable, and that some are more involved in horizontal transfer between genera than others. This suggests the possibility of developing data-driven methods for statistically compensating for the contact signal and amplifying the signal of phylogeny. A spatial analysis of ‘‘typological distance’’ (presented section 3) gives a measure of the influence of horizontal, non-genealogical transfer in Island Melanesia.

A detailed description and rationale for the database used in the present analysis is given in Dunn et al. (2005, including Online Supplementary Materials); see Appendix A. In brief, 124 structural features were encoded for 16 Austronesian languages from the Western Oceanic subgroup and 15 Island Melanesian Papuan languages from the same area. These features were all abstract binary characters, i.e., presence or absence of a grammatical feature without reference to formal expression.

The investigation of structural influence between Oceanic and Papuan languages in Island Melanesia is only possible because the Oceanic languages have an independently known phylogeny. A crucial logical step to this investigation is to demonstrate (as will be done in sections 2.1 and 2.2) that structural linguistic features can carry a phylogenetic signal. Once it has been shown that the computational method for tree generation can replicate the results of the comparative method for the Oceanic languages, then a genealogical tree on the Papuan languages generated using the same methods should be seriously considered as a hypothesis. The results of the uncontrolled study need careful interpretation, particularly with respect to their intrinsic plausibility. In the subsequent section (section 3) I will

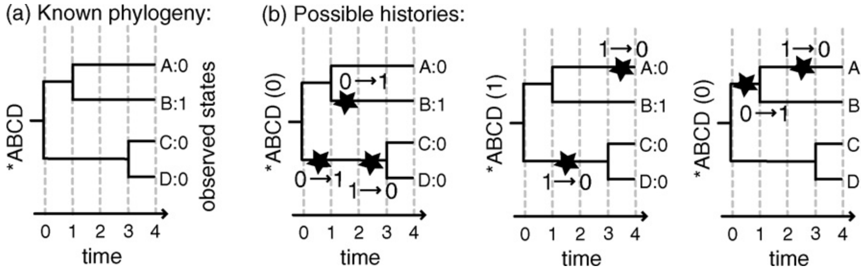

Fig. 2. Possible histories of the values of a trait from a simulation. Stars on branches mark points where the character state flips from one state to the other.

discuss methods for evaluating the Papuan tree as a phylogenetic hypothesis, and will test some counter hypotheses, which explain the structure of the Papuan tree from other, non-genealogical, factors, such as structural convergence. It will be shown that the set of relationships shown by the linguistic structural data is plausible from a geographic perspective, and that non-phylogenetic accounts of the geographic pattern fit the data poorly. This analysis has also proved useful in the generation of testable hypotheses, particularly in the realm of human genetics (e.g., Cox and Lahr, 2006; Hunley et al., 2007).

2.1. MCMC Bayesian phylogenetic analysis

Dunn et al. (2005) study of this data used the *maximum parsimony* method for inferring phylogeny. Maximum parsimony is an intuitively clear method, which generates the tree requiring the fewest state changes (i.e., changes in the values of each feature) to evolve from an ancestor to all the observed descendants. The parsimony method has been criticised for not having a ‘‘model’’ of evolution, and for not being statistical.[1] This makes it difficult to interpret the results of a parsimony analysis: unrealistic assumptions may limit the appropriateness of the analysis, and there is no statement of statistical confidence. Phylogenetic inference from a probabilistic, model based perspective do not have these shortcomings. Model based methods maximize *likelihood*, that is, they infer the history most likely to produce an outcome the same as the observed state of affairs. The prerequisites for a likelihood calculation are (i) a *model* of the evolutionary process, (ii) a *tree topology*, and (iii) specified *parameter values* for that model. In these analyses, the model is specified by the investigator (chosen empirically, or from a priori knowledge of real evolutionary processes); the tree topology and parameter values are the object of the search.

The phylogenetic model describes the behaviour of individual character traits. An extremely simple (and unrealistic) model could be that ‘‘each character changes state with probability *P* per unit time’’. Given observations of the states of a particular character shown in the hypothetical phylogeny in panel (a) of Fig. 2, character history can be simulated by randomly (with probability *P*) flipping the character state moving along the tree from the root to the tips. Panel (b) shows some possible character histories that give the same ultimate distribution character states as observed in the ‘‘real’’ distribution. The simulation would also produce many result distributions which were different to the observed distribution. If the simulation was carried out a large number of times for a particular value of *P*, the proportion of result distributions equivalent to the observed distribution would be analogous to the likelihood of that particular value of *P* (for that model and tree). If the simulation was carried out for a well-distributed sample of all the possible values of *P*, the (approximate) *P* value with *maximum likelihood* of producing the observed distribution could be identified. Thus, the best phylogenetic explanation of the data for a given model is the combination of tree topology and parameter values which has the highest likelihood of producing the observed variation.

Of course real data involves many more than the one character in the toy example above, and the single-parameter model used in this example is clearly inadequate. More realistic models might ‘‘fit’’ the data better (again in the sense of being more likely to produce the equivalent distribution in a simulation). Additional parameters that might be added to a model include allowing for several different rate classes of character (so that some characters are allocated to conservative rate classes, where character states change infrequently on branches and sisters within branches tend to have the same values, and some characters are more innovative, giving lots of variation). Another realistic parameter

[1] The bootstrap procedure is partial a work-around for the lack of statistical evaluation; Tuffey and Steel (1997) shows that parsimony and model-based ‘‘likelihood’’ methods retrieve the same tree under certain models.

that might be added to a model is to distinguish ‘‘gains’’ from ‘‘losses’’, that is, to allocate characters to separate rate classes for changes from 0 → 1 as from 1 → 0. This is particularly appropriate for traits representing reflexes of lexical cognate sets (as used by, e.g., Gray and Jordan, 2000; Gray and Atkinson, 2003; Greenhill and Gray, 2005), as these are very asymmetrical: highly unlikely to be spontaneously innovated, but relatively easily lost. This kind of parameter is not appropriate for use when searching for unrooted trees however, since it is not possible to specify what counts as a gain and what counts as a loss in a tree that doesn’t specify which nodes are ancestral to which; in the phylogenetic analysis of a group of languages not known to be related it may be difficult to justify a particular rooting hypothesis.

The technique of determining phylogeny by calculating the tree topology and parameter values which maximize the likelihood of the observed data is called, unsurprisingly, Maximum Likelihood. But direct calculation of likelihood is computationally difficult, and may be computationally intractable for large amounts of data. Bayesian Phylogenetic Inference provides an effective heuristic for coping with such large data sets.[2] It is an iterative, stochastic process, which searches the vast ‘‘space’’ of possible trees and model parameter settings and returning a sample of those with the maximum likelihood of accounting for the observed data. This space is huge: for the Oceanic language sample of 16 taxa there are no fewer than 213 458 046 676 875 distinct unrooted tree topologies alone (Felsenstein, 2004:24); and the likelihood of each of these tree topologies varies continuously with parameter settings of the model. The type of Bayesian Phylogenetic Inference described here is known as MCMCMC the three MCs refer to three key elements of the method, ‘‘Monte Carlo’’, ‘‘Markov Chain’’ and ‘‘Metropolis Coupled’’. These will be described below.

Bayesian inference is a statistical technique which quantifies *conditional probability,* that is, how much the estimate of the probability of a hypothesis being true should be revised based on further observations. The initial estimate is called the *prior probability* and the revised estimate is called the *posterior probability.* In Bayesian Phylogenetic Inference, the conditional probability calculation is carried out between two points in the parameter space, to determine which one is more likely. The result of this comparison functions as a new prior, which in turn allows a new posterior probability to be estimated in an iterative search.

The *Monte Carlo Markov chain* searches the parameter space following a conceptually simple algorithm in which the likelihood is compared between the current position in the space and another randomly chosen (hence *Monte Carlo*) position nearby. If the new position in the parameter space has higher likelihood than the current one, it becomes the current position for the next iteration of the search (these repeated searches form a *Markov chain*, a procedure which retains no memory of its previous states). As long as the likelihood of the new position in the parameter space is higher than the current position then this algorithm is unproblematic. But if there are local optima, areas in the parameter space which have higher likelihood than neighboring areas, but which are not the overall optimal likelihood zones, then this kind of search can get stuck. *Metropolis coupling* is a refinement of the algorithm to allow the search to escape local optima: if the newly sampled likelihood is lower in an iteration of the Markov chain, a random choice is made between keeping the priors from the current round or using the posteriors, with a probability of choosing the latter equal to the ratio of the new to old likelihood scores. The use of Monte Carlo Markov chains to search the parameter space is schematically represented in Fig. 3. The Markov chain does not retain any memory of previous states beyond the most recent pair being compared. Each iteration moves stochastically through the parameter space, tending towards areas of greater likelihood, until it reaches the equilibrium zone, an area of the parameter space with consistently higher likelihood. The goal of the search is a sample of trees/parameters from the equilibrium zone. Before this state is reached the likelihoods fluctuate wildly with each iteration (likelihood values over a search are plotted in Fig. 4), and the trees produced do not mean anything much—this is called the burn-in period, and these trees are discarded from the analysis.

Given a sufficiently strong phylogenetic signal, the starting point of the search doesn’t matter, but for a weak signal the choice of priors will influence the outcome of the analysis. There are good arguments for starting the search somewhere near the area of the parameter space that the experimenter knows must contain the answer, but this is of course only possible where the experimenter has a priori knowledge of the probable distribution of parameters.

As stated above, it is possible to decide empirically which model is best: a phylogenetic inference can be made using a range of different models, involving different sets of parameters, and the model which produces the best (highest likelihood) results can be identified. Fig. 4 shows the likelihood scores of 12 million iterations of tree

[2] Empirical tests show that Bayesian methods usually perform better than parsimony: they are more likely to retrieve a phylogenetic signal present in the data, and allow a greater degree of confidence in the results obtained (Ronquist, 2004)(see also Holden et al., 2005 for a comparison of parsimony and Bayesian methods using cultural data).

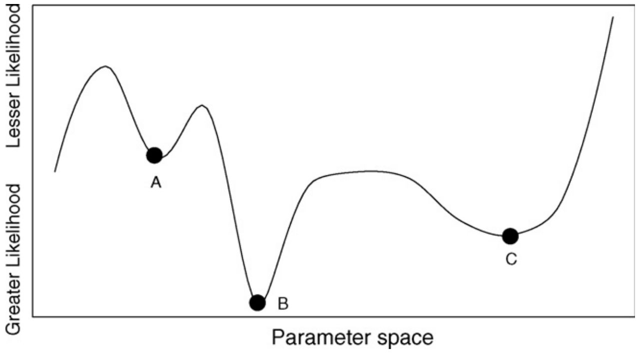

Fig. 3. Search for the equilibrium zone (greatest likelihood, represented as the lowest point) in a one-dimensional parameter space (represented as a position along the *x*-axis). Points A and C are local minima, but point B is the global minimum. The search can start from any point on the line; and follows a pattern tending ever downwards, dropping easily into a local equilibrium. Metropolis coupling allows occasional upwards moves (more likely when the likelihood ratio is small, less likely when it is big). Simulated annealing (described Press, 1988:444) is another technique to help the search avoid getting stuck in non-optimal, local minima. The search is ''heated'' (search parameters are allowed to vary randomly), and then slowly ''cooled'' so that the amount of random variation is gradually reduced. The physical analogy would be trying to roll a ball-bearing to the bottom of an irregularly dented metal bowl. At the start of the search the bowl is shaken vigorously (the ''heating''), and then slowly brought to rest.

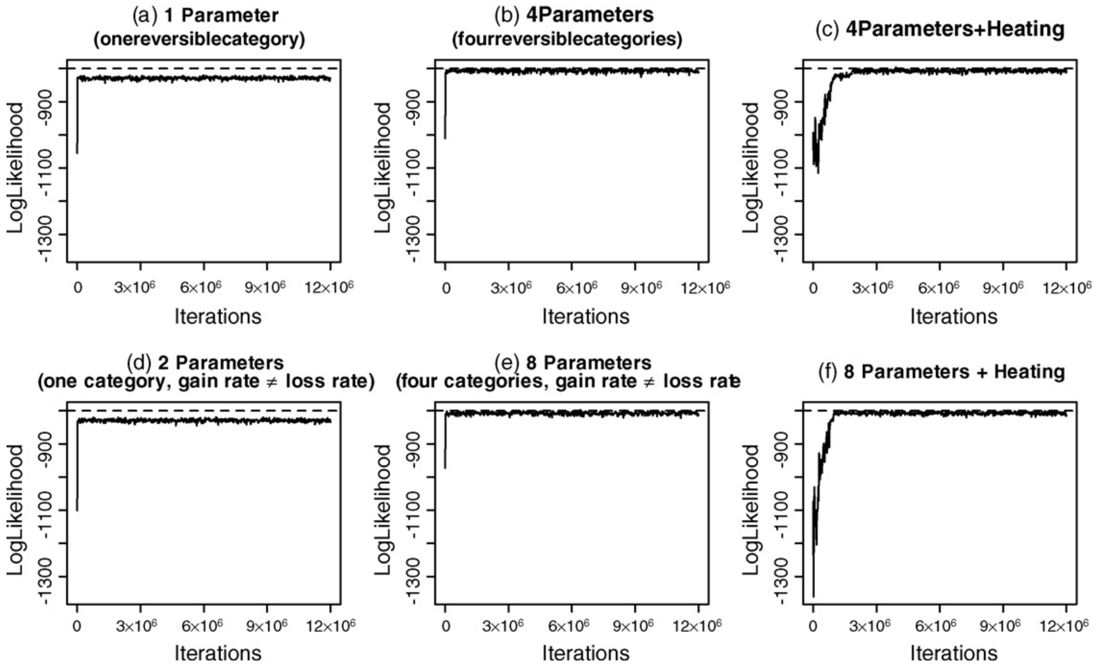

Fig. 4. Equilibrium of different models. Models (a)–(c) have a single rate for gain and loss, models (d)–(f) have a separate rate for each. The ''one rate'' models use the same rate category for all features; the ''four rate'' models divide features into four different rate categories; Models (c) and (f) are ''heated'' for the first 1 500 000 iterations. With this data, model (b) behaves better than (a) and (d) – the models with fewer parameters – and not appreciably worse than the other, more complex, models.

searching under six different models for the Oceanic language data. In all cases they achieve equilibrium fairly quickly (within 1–3 million iterations), but the equilibrium level varies.[3] When modeling the behavior of typological features a fairly simple model performs as well as any. This model is *reversible*, meaning that gain and loss probabilities are not

[3] Bayesian phylogenetic analyses produce poor results when the model is misspecified. Experiments with simulated data show that this is a particular risk with an underspecified (i.e., overly simplistic) model—a model which specifies more parameters than necessary (i.e., the simulation specified more parameters than were used to generate the data) performs roughly as well as a well-specified model (Huelsenbeck and Rannala, 2004). But it is wise to keep the model as simple as possible: too complex a model (technically, an *over fitted* model), while explaining the observed data well, may not generalize to new data.

calculated separately (this also makes it possible to remain agnostic about the chronological ordering of diversification and infer an unrooted tree—otherwise, to determine what constitutes a gain and what a loss the tree must be directional, and so it is necessary to decide a priori upon a root). More importantly, this model requires typological features to have different rate categories (the experimenter only determines that these rate categories will exist: the allocation of features to rate categories is part of the Bayesian inference process). Having a set of different rate categories allows there to be some features which are more stable within lineages, and some which are more variable. It is thus part of the analysis that some features are not strongly diagnostic of subgroups; and this is only to be expected for typological features, since these features constitute a relatively small design space, and there is high chance of accidental convergence (homoplasy). But the identity of the more and less conservative features for the phylogeny emerge from the analysis, they are not chosen by the experimenter in advance.

Once in the optimal zone the function is in equilibrium (since deviation from these optimal parameters result in lower likelihoods). The trees produced during the equilibrium state of the Markov chain constitute a sample of equally good phylogenetic hypotheses. Once a suitably large sample of trees has been generated, the information these trees contain can be summarised using, e.g., a consensus tree or a consensus network (methods outlined in section 2.2).

2.2. Bayesian Phylogenetic Inference and the Oceanic languages

As stated above, the results of the Bayesian phylogenetic analysis are a set of more-or-less equiprobable trees. No one particular tree in this set is the ‘‘correct’’ tree. One way to summarise the information contained in this tree set is to use some kind of consensus representation. Traditionally this is done with a *c* onsensus tree, a tree which is built by transcribing all the non-conflicting branchings present in the tree set in order of frequency, i.e., starting from a completely unresolved, star-like consensus tree, split it according to the most commonly occurring bifurcation the tree set; working down in frequency order through the tree set add all further bifurcations that don’t contradict one already added, until the consensus tree is completely resolved, or until there are no more branchings present in more than 50% of the trees. This is illustrated in Fig. 5. Panel (i) shows a toy tree set, with only three members (matching splits between the trees are marked here and in the other panels). Panel (ii) lists the bifurcations present in the tree set, and (iii) shows the consensus tree of this data. The numbers on the branches indicate the percentage of trees in the tree set with the branching. A more sophisticated version of this, the *consensus network* (Bryant and Moulton, 2004), illustrated in (iv), preserves some of the conflict present in the tree set as well, giving what is probably the best single representation of the results of a Bayesian phylogenetic analysis. The parallel lines show the presence of a split in the data, with length proportional to the frequency the split is attested. A parallelogram in this kind of graph shows that there is conflicting evidence, supporting incompatible bifurcations in the tree.

Figs. 6 and 7 show the results of a Bayesian Phylogenetic Inference carried out on the Oceanic data. The consensus tree representation of the posterior tree sample of the Oceanic languages is not very well resolved, with only the

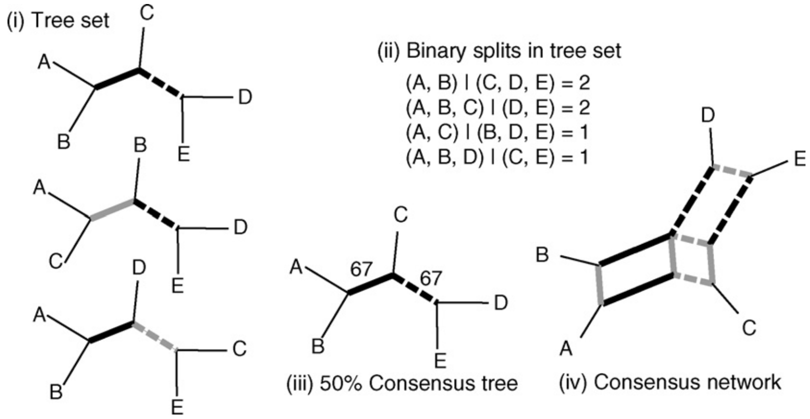

Fig. 5. Representing consensus in a set of trees. (i) Sample tree set containing three trees; (ii) tree bifurcations in the tree set, by frequency; (iii) consensus tree, with branch frequency marked; (iv) consensus network of tree set (splits are shown by parallel lines; the length of these lines is proportional to the number of times the split is observed in the data).

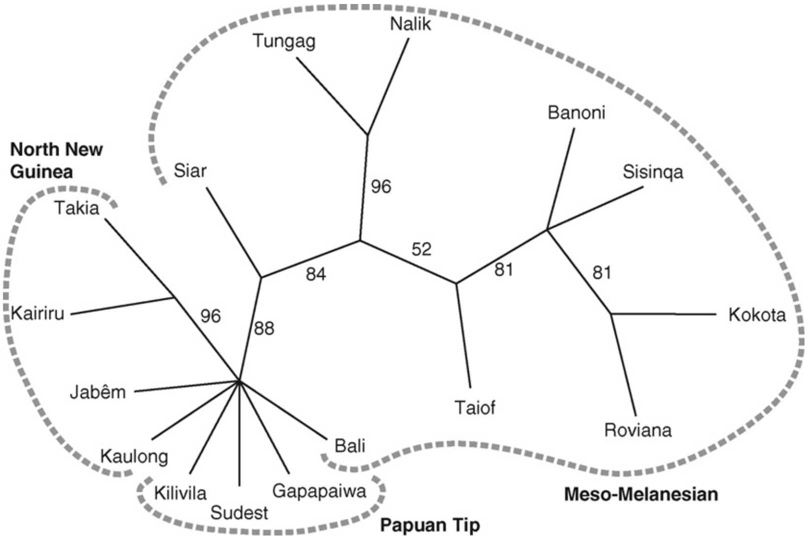

Fig. 6. Consensus tree of the resulting tree set from Bayesian phylogenetic analysis (numbers on branches show the percentage of trees with this branching; dashed lines mark affiliation by comparative method).

Meso-Melanesian subgroup showing significant structure (Fig. 6). Most of the members of the Meso-Melanesian subgroup of the tree are correctly placed with respect to each other. The exceptions are (i) Bali, which ought to be the earliest branching in the Meso-Melanesian clade, and (ii) the branching order of Siar and the Tungag-Nalik clade, which should be reversed (see the comparative method classification, Fig. 8). However, the North New Guinea and Papuan Tip clades are not resolved at all beyond the (correct) clustering of Kairiru and Takia. In contrast, the consensus network representation of the tree sample is in nearly perfect agreement with the comparative method tree (Fig. 7). Some conflict in the tree is visible which pulls Sudest away from its comparative method sisters Gapapaiwa and Kilivila of the Papuan Tip cluster, and causes it to move somewhat towards the North New Guinea languages. This, and a little bit of conflict within the North New Guinea languages (perhaps a result of conflict pulling Kaulong toward the other languages of the island of New Britain) is enough to cause the consensus tree to lose almost all resolution for these groups.

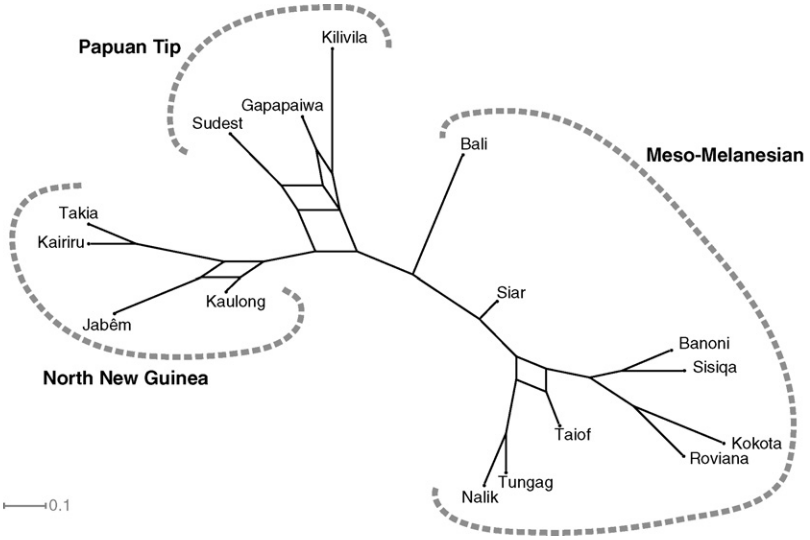

Fig. 7. Consensus network of Bayesian trees of the Oceanic sample (dashed lines mark affiliation).

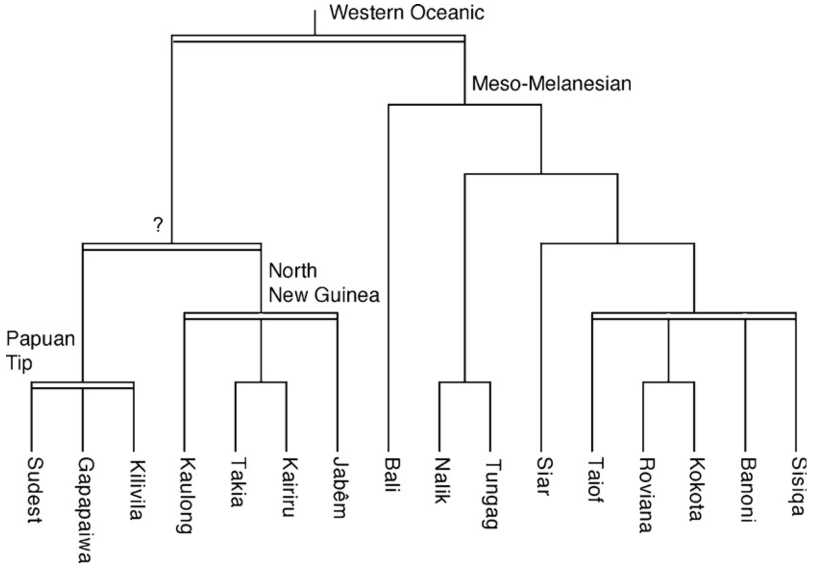

Fig. 8. The comparative method tree of the Western Oceanic languages in the sample (double lines are linkages; the node marked with the question mark is uncertain: Papuan Tip and North New Guinea may emerge directly from Western Oceanic; Lynch et al., 2002:101).

Subjective impressions of the similarity/difference between trees may be insufficient, in which case quantitative measures are required for comparing the similarities of two trees. A reasonably sensitive, yet computationally tractable, measure of tree-to-tree distance is the *quartets distance* (Felsenstein, 2004:530). The quartets distance counts the number of four-member subtrees ('quartets') shared between the two trees being compared. Given a known reference tree, it is possible to measure the distance of a tree inferred in a phylogenetic analysis to this standard.[4] In this study, the quartets distance was calculated using the program QuartetDist, a tool implementing a method for measuring quartet distance which allows for non-binary trees (Christiansen et al., 2005, 2006).

The selection of the precise tree to measure also poses a problem. One possibility would be to measure the distance from the consensus tree to the reference tree. This has the advantage that there is just one consensus tree, so that that there is a single measure of similarity. Another possibility is to measure the distance to the reference tree from each of the equiprobable trees generated in the Bayesian analysis. The set of distances can then be represented as a histogram. This has the advantage that better fit by subsets of this tree set will be obvious.

Given a measure of the difference between the analytic results and the reference tree, it is possible to evaluate whether this degree of similarity is likely to have occurred by chance. This can be done by comparing the set of distances from the result trees to the reference tree, to the distances from the reference tree to a set of random trees. Fig. 9 shows superimposed histograms with distances measured from the reference tree to a set of randomly generated trees and from the reference tree to the result set of equiprobable trees from the Bayesian analysis. Very few of the random trees are better (i.e., shorter) than even the worst (longest) of the Bayesian trees. The quartet distance of the consensus tree to the reference tree is marked with the vertical line. Note that none of the 1000 randomly generated trees are as good as this one.[5]

The Bayesian method of phylogenetic inference can, in the case of Oceanic languages, retrieve a set of trees which are highly congruent with the comparative method tree, despite being based on completely independent types of data.

[4] The quartets distance is superior to the commonly used *partition metric*. The partition metric has the advantage of computational simplicity, and is implemented in commonly available phylogenetic software such as Swofford (2003), but has a major disadvantage in that it can report trees as maximally distinct which differ in the placement of a single taxon (Penny and Hendy, 1985).

[5] Note that measurements of similarity or difference between two sets of trees have to be taken with a grain of salt. Evaluation of the similarity of trees derived by an analysis to a known reference tree is known to be problematic (Felsenstein, 2004:534). For example, if one pair of taxa in the data are so similar that they are always sisters in the analysis trees, this may be enough to make the set of analytic trees significantly more similar to the reference tree than the random trees.

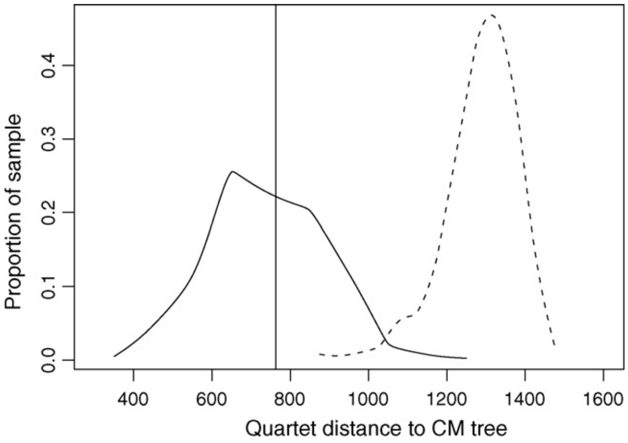

Fig. 9. Histograms of quartet distances from the comparative method tree to the Bayesian tree sample (solid) and a set of randomly generated trees (dashed). The vertical line shows the quartet distance of the CM tree to the Bayesian consensus tree.

3. Genealogical and contact signal: Papuan languages

The Papuan data was analyzed using the same Bayesian methods as described above. In this case there is no independently established reference tree. It must be kept in mind that tree inference methods *always* produce a tree; it cannot be that the method refuses to attach one or more taxa to the tree. Once the best tree (or, as in this kind of analysis, the set of best trees) is generated, it is the task of the analyst to decide whether the apparent relationships shown by the tree are motivated by chance, or represent some real historical process.

The corresponding consensus network for the Papuan languages is given in Fig. 10. As in the results of the parsimony analysis reported previously in Dunn et al. (2005), this network shows Bougainville languages clustered at one end, Bismarcks languages clustered at the other, and the Central Solomons languages positioned between them.

Note that there is a considerable amount of conflict present in this tree. In particular, there are trees clustering Lavukaleve with Bougainville/Central Solomons (congruent with geography), but also with some or all of the Bismarck Archipelago languages. This is consistent with a hypothesis of deep historical relations between the Papuan languages of the Solomons and Bismarcks.

There is no independent historical linguistic standard for the possible relationships between the Papuan languages of Island Melanesia. This network is however highly plausible, as it contains regions which correlate well with the geographic distribution of these languages over archipelagos. A number of methods exist with which to investigate this

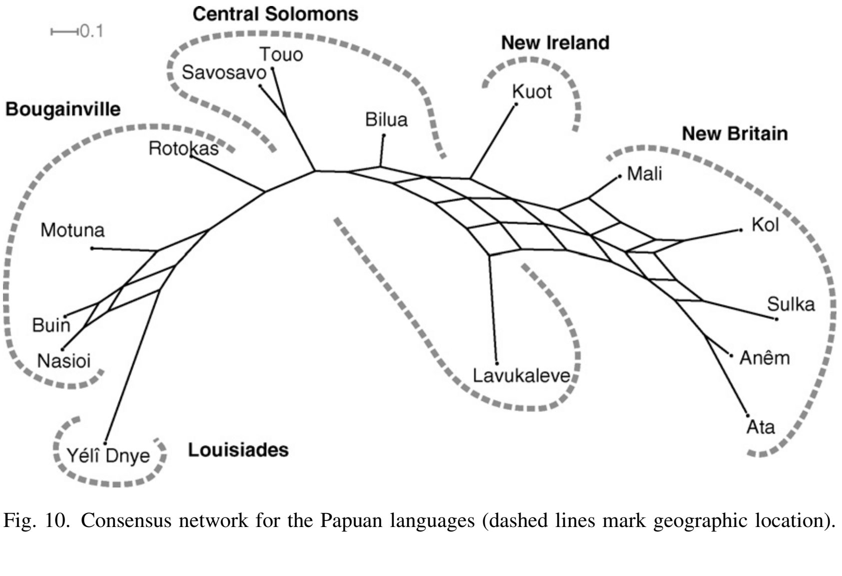

Fig. 10. Consensus network for the Papuan languages (dashed lines mark geographic location).

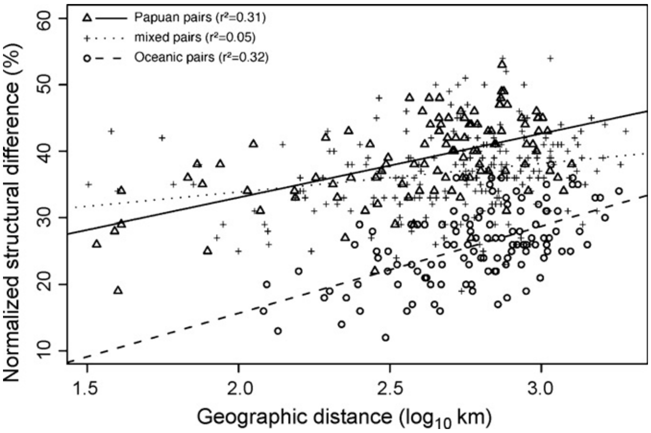

Fig. 11. Spatial correlation with typological structure for Oceanic (circles), Papuan (triangles) and Mixed (dots) pairs of languages.

geographic patterning. A correlation between structural dissimilarity and spatial distance – so that geographically nearer pairs of languages tend to be structurally more similar, and geographically more distant languages tend to be structurally more different – provides evidence of some sort of historical interaction. If this correlation obtains across the Oceanic/Papuan divide then there would be firm evidence that non-phylogenetic processes are responsible. But if spatial–structural correlations are strongest between languages from a single group, then this is evidence for a phylogenetic relationship, since contact influence should not be sensitive to genealogy.

Structural dissimilarity between two languages is measured using normalized Hamming distance (percentage difference between all pairs of values of typological features, excluding features where one or both of the languages have ‘‘unknown’’ value). Geographic distance is measured a logarithmic scale, so that less note is taken of differences between large distances than between small distances. A matrix of these pairwise structural distances can be compared statistically to a matrix of pairwise geographic distances to see to what degree they are correlated. Using a simple linear model, the correlation between structural and log geographic distances for all language pairs in the sample is small but clear, with $r^2 = 0.03,\ p = 0.0001$, i.e., geographic distance predicts 3% of the variance in the structural data.

Fig. 11 gives a pairwise comparison of linguistic differences (percentage disagreement between the coded data for each pair of languages) plotted against log geographic separation for the language pairs divided into three groups, Oceanic–Oceanic, Papuan–Papuan, and mixed pairs of one Oceanic, one Papuan language.[6] This figure shows that, while the structural distance between pairs of Papuan languages is in general greater than the structural distance between pairs of Oceanic languages, nevertheless structural distance accumulates for Papuan languages with increasing geographical distance (see also Lindström et al., 2007). The r^2 values for Oceanic–Oceanic and Papuan–Papuan pairs are quite high: more than 30% of the variance in structural difference can be predicted from geographic separation. But the correlation between mixed pairs of languages is low, with $r^2 = 0.05$, showing that, e.g., a Papuan language is unlikely to be much more similar to nearby Oceanic languages than distant ones.

There are a few obvious accounts for the structured similarity relationships between the Papuan languages of Island Melanesia. Chance can be ruled out, since the correlation with geography is much too good. Contact between Papuan languages could hypothetically lead to structural convergence with neighbours, thereby creating a geographic pattern in genealogically unrelated languages. This can probably be ruled out too: contact between Papuan languages seems unlikely as a structuring principle in the geographic correlation given the low density of Papuan languages in Island

[6] An anonymous reviewer expressed doubts that distance ‘‘as the crow flies’’ is an appropriate measure of separation. In general this is true, but with the particular language sample we have in Island Melanesia it is a reasonable approximation. The geographic analysis has been carried out measuring distance via waypoints positioned between each major island group to approximate distance ‘‘as the canoe paddles’’. Because of the largely linear layout of the islands in the archipelago from the Bismarcks to the Solomons (see Fig. 1), the correlation between crow- and canoe-distances is very high ($r^2 = 0.78$ for distance measured on the log scale). The use of distance as the crow flies consistently underestimates distance to the four languages Louisiades, but has little effect on any of the other pairs. The better model of human accessibility provided by the use of waypoints increases the correlation between structural and geographic distance for the two homogeneous groups (Oceanic–Oceanic and Papuan–Papuan pairs), and decreases it for the heterogeneous (Papuan–Oceanic pairs).

Melanesia (20 languages out of 200), and that the Papuan languages are mostly relatively isolated from each other by distance and by intervening Oceanic languages. The most strongly clustered pairs in the Papuan consensus tree (not shown) are Touo and Savosavo, which are separated by about 250 km, including a 80 km sea crossing, and Anêm and Ata, which are separated by more than 200 km—much of which is uninhabitable mountains (however, in the case of Anêm and Ata and the other New Britain languages, there is a known history of population movement due to volcanism). In both cases there are a great number of Oceanic languages in closer proximity. It *is* possible that these languages are convergent due to Papuan on Papuan contact effects from a period prior to the expansion of the Oceanic languages. However, this would put the period of interaction back several thousand years, which makes it equally interesting to linguistic prehistory as descent from a common ancestor would be.

This leaves the most likely explanation for the structural-spatial correlation between Papuan languages to be shared patterns of Oceanic contact, or common ancestry.

There certainly are contact effects between Papuan and Oceanic languages. In some cases Papuan languages share features which are typical of Oceanic languages, making them likely candidates for borrowing into Papuan from Oceanic. In other cases subgroups of Oceanic languages have distinctly non-Oceanic features, and it has been supposed that these features are the result of contact from Papuan languages. Many of the features so treated are common in the Papuan languages of Island Melanesia (e.g., verb-final word order). The scatter plot in Fig. 11 shows a slight increase in structural distance between mixed Papuan–Oceanic pairs of languages, which is further evidence of the inter-genus contact effect. However both the Papuan languages and the Oceanic languages show much stronger patterns of geographic correlation in their patterns of in-group similarity, and a much weaker correlation to geography between members of the mixed Papuan–Oceanic pairs. Note also that the Papuan languages do not currently have closer contact relations within the Papuan group than between Papuan and Oceanic (Oceanic languages may be different in this regard—the relatively dense and homogeneous distribution of Oceanic homogeneity may facilitate linguistic contact effects with other Oceanic languages, rather than with the sparse and diverse Papuan languages). If the geographic patterning in the Papuan languages was caused by contact effects, one would not expect proximity effects of Papuan–Papuan pairs to be stronger than proximity effects of Papuan–Oceanic pairs.

Current contact is thus very unlikely to explain the structured relationships between the Papuan languages of Island Melanesia, which lends strength to the hypothesis that the network in Fig. 10 represents a phylogenetic (or at least pre-Austronesian) signal.

4. Conclusion

The statistical analysis of the tree sets produced for the Oceanic and Papuan groups of languages has shown that within each of the Papuan and Oceanic groups, languages have spatially correlated structural variation consistent with separate phylogenetic processes. There is also weaker evidence of spatial correlation between Papuan and Oceanic languages, consistent with structural convergence after initial independent diversification. However, the results are not consistent with a pattern of later convergence.

Thus, two hypotheses – not necessarily mutually exclusive – remain as most likely: either the relationships between the Island Melanesia Papuan languages are due to common ancestry, or these relationships are due to ancient convergence, prior to the arrival of the Oceanic languages. In either case, what has been detected is a signal of deep linguistic prehistory (confirming the conclusion of Dunn et al., 2005). The real world significance is also similar. If the signal is due to convergence, it points to extremely intensive interaction between the ancestors of these languages, since this structural convergence is still detectable after several thousand years of separation. Common ancestry is a simpler explanation.

In summary, the structural similarities between the Papuan languages of Island Melanesia are:

- *not* motivated by Oceanic contact
- *not* motivated by recent Papuan contact
- *perhaps* motivated by ancient Papuan contact, but
- *most likely* the result of an ancient phylogenetic signal

Looking from a broader perspective than simply the issues in Island Melanesian Papuan linguistics, it is worth noting that these results show that abstract structural features of language can carry a phylogenetic signal (as shown quite conclusively by the Oceanic data), but also that it may be the case that linguistic typology can continue to carry a detectable phylogenetic signal after the lexical signal of common origin has decayed. The results of this study support the earlier results of Dunn et al. (2005) to suggest that the Island Melanesian Papuan languages descend from a common ancestor, despite the lack of reconstructible lexical cognates. These results do not show that one should expect greater conservatism in language structure than in lexicon in all cases; but potential counterexamples, where lexicon preserves a signal absent from language structure, would not make the instances of ‘‘conservative structure’’ less interesting or important.

In 2005 there was hope that (biological) genetics would provide the ultimate solution to these problems. Recent publications on language and gene correlations in Island Melanesia given cause to moderate this optimism: gene flow between populations seems to obscure the genetic signals of the ancestral language groups over the time scale we are dealing with (Cox and Lahr, 2006; Hunley et al., 2007). It is still conceivable that lexical evidence will be able to provide further demonstration of historical links between the non-Austronesian languages of Island Melanesia: Ross' work with pronouns provides a beginning (Ross, 2001), and a composite approach to reconstruction using the comparative method along with statistical judgements of stability (see, for example, Pagel et al., 2007) or probabilistic cognacy judgements may yet give results.

Adherence to the comparative method has provided rigor to counterbalance other more dramatic but ultimately unverifiable methods (e.g., Greenbergian approaches and the other ‘‘megalo-comparativists’’ (Matisoff, 1990); But computational phylogenetic and areal/geo-statistical approaches allow us to investigate linguistic change outside the scope of the comparative method (as has been shown earlier, e.g., Nichols, 1992). These approaches promise possibilities to get beyond the comparative method in several ways. They offer a way of demonstrating monophyly (common ancestry) for languages where there is no detectable lexical signal. They also provide tools which allow the traditional parthenogenetic assumptions (i.e., that a language has a single parent) to be relaxed.

Acknowledgements

This paper is a reanalysis of data published in Dunn et al. (2005), and I am indebted to my colleagues Robert Foley, Eva Lindström, Stephen Levinson, Ger Reesink, Gunter Senft, Angela Terrill from the ‘‘Pioneers of Island Melanesia’’ project for the development of that project, particularly the design of the questionnaire and for coding of typological data from fieldwork and descriptions; Stuart Robinson, Claudia Wegener also provided typological data from fieldwork, and Victoria Rodriguez, Sebastian Nordhoff coded data from written descriptions. Thanks also to: Niclas Burenhult, Simon Greenhill, Eva Lindström, Ger Reesink, Angela Terrill, Claudia Wegener for general comments and criticism of this paper, as well as to Russell Gray, John Nerbonne and two anonymous reviewers; Sean Myles and Dan Dediu gave advice on investigating spatial correlation; Harald Baayen and Jan-Peter de Ruiter for help with statistics; Søren Wichmann for pointing out the need to test tree congruence statistically; Russell Gray and Simon Greenhill for technical advice and help with the Bayesian phylogenetic analysis; This paper was originally presented at the special session ‘‘Comparing Aggregate Syntaxes’’ at the Digital Humanities conference in Paris on July 6, 2006, organized by John Nerbonne and Franz Manni. The ‘‘Pioneers’’ database, described in Dunn et al. (2005), was developed as part of the European Science Foundation EUROCORES Programme OMLL, and was supported by funds from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO); the Arts and Humanities Research Council, UK; and the EC Sixth Framework Programme under contract no. ERAS-CT-2003-980409. Current support is acknowledged from NWO project ‘‘Breaking the time barrier: Structural traces of the Sahul past’’ (P.I. Pieter Muysken), Radboud University Nijmegen, and I am grateful also for additional support from the Max Planck Institute for Psycholinguistics, Nijmegen.

Appendix A. Database

The language sample and typological questionnaire contents used in this paper are purposely unchanged from Dunn et al. (2005) study, since the intent of this paper is in part to show that the improved methods described produce better results than were obtained in the original study. The data (including description of feature definitions and sources) is available at http://www.sciencemag.org/cgi/data/309/5743/2072/DC1/1. Further discussion of the rationale for the selection of languages and typological features is presented in Dunn et al. (2007). The original team has also made improvements to the language sample and to the questionnaire: these are described in Dunn et al. (2008).

The Oceanic languages mentioned in this paper can be identified by their ISO-639-3 language codes (see Gordon, 2005:7) as: Bali (bbn), Banoni (bcm), Gapapaiwa (pwg), Jabêm (jae), Kairiru (kxa), Kaulong (pss), Kilivila (kij), Kokota (kkk), Nalik (nal), Roviana (rug), Siar (sjr), Sisiqa (qss), Sudest (tgo), Taiof (qtf), Takia (tbc), Tungag (lcm). The Papuan languages are: Anêm (anz), Ata (ata), Bilua (blb), Buin (buo), Kol (kol), Kuot (kto), Lavukaleve (lvk), Mali (gcc), Motuna (siw), Nasioi (nas), Rotokas (roo), Savosavo (svs), Sulka (sua), Touo (tqu), Yélî Dnye (yle).

Appendix B. Technical note

The key statistical and phylogenetic tools used in the analyses reported in this paper were: APE (an R library providing tools for phylogenetics and tree drawing; Paradis et al., 2003); BayesPhylogenies (Bayesian Phylogenetic Inference; Pagel and Meade, 2004); QuartetDist (quartet distance measurement; Christiansen et al., 2005, 2006); SplitsTree4 (consensus network; Huson and Bryant, 2006); PAUP (consensus trees and random tree generation; Swofford, 2003). The R statistical language was used for general statistics and plotting (www.r-project.org).

References

Adelaar, A., Himmelmann, N., 2005. The Austronesian Languages of Asia and Madagascar. Routledge, London.

Bellwood, P., Fox, J., Tryon, D., 1995. The Austronesians: Historical and Comparative Perspectives. Department of Anthropology, Research School of Pacific and Asian Studies, Canberra.

Bryant, D., Moulton, V., 2004. NeighborNet: an agglomerative algorithm for the construction of planar phylogenetic networks. Molecular Biology and Evolution 21, 255–265.

Bryant, D., Filimon, F., Gray, R., 2005. Untangling our past: languages, trees, splits and networks. In: Mace, R., Holden, C., Shennan, S. (Eds.), The Evolution of Cultural Diversity: A Phylogenetic Approach. UCL Press, London, pp. 67–84.

Christiansen, C., Mailund, T., Pedersen, C., Randers, M., 2005. Computing the quartet distance between trees of arbitrary degrees. In: Proceedings of Workshop on Algorithms in Bioinformatics. Springer-Verlag, New York, pp. 77–88.

Christiansen, C., Mailund, T., Pedersen, C., Randers, M., 2006. Tools for calculating the split- and quartet-distance for sets of trees of arbitrary degrees. In: Christiansen, C., Randers, M. (Eds.), Computing the Quartet Distance Between Trees of Arbitrary Degrees. University of Aarhus, Department of Computer Science, pp. 89–106.

Cox, M., Lahr, M., 2006. Y-chromosome diversity is inversely associated with language affiliation in paired Austronesian-and Papuan-speaking communities from Solomon Islands. American Journal of Human Biology 18, 35–50.

Dunn, M., Foley, R., Levinson, S., Reesink, G., Terrill, A., 2007. Statistical reasoning in the evaluation of typological diversity in Island Melanesia. Oceanic Linguistics 46, 388–403.

Dunn, M., Levinson, S., Lindström, E., Reesink, G., Terrill, A., 2008. Structural phylogeny in historical linguistics: methodological explorations applied in Island Melanesia. Language 84 (4), 710–759.

Dunn, M., Reesink, G., Terrill, A., 2002. The East Papuan languages: a preliminary typological appraisal. Oceanic Linguistics 41, 28–62.

Dunn, M., Terrill, A., Reesink, G., Foley, R., Levinson, S., 2005. Structural phylogenetics and the reconstruction of ancient language history. Science 309, 2072–2075.

Felsenstein, J., 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, Massachusetts.

Gordon, R., 2005. Ethnologue: Languages of the World, 15th ed. SIL International, Dallas, TX.

Gray, R., Atkinson, Q., 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature 426, 435–439.

Gray, R., Jordan, F., 2000. Language trees support the express-train sequence of Austronesian expansion. Nature 405, 1052–1055.

Greenberg, J., 1971. The Indo-Pacific hypothesis. In: Sebeok, T. (Ed.), Current Trends in Linguistics, Vol. 8: Linguistics in Oceania. Mouton and Co., The Hague, pp. 807–871.

Greenhill, S., Gray, R., 2005. Testing population dispersal hypotheses: pacific settlement, phylogenetic trees and Austronesian languages. In: Mace, R., Holden, C., Shennan, S. (Eds.), The Evolution of Cultural Diversity: A Phylogenetic Approach. UCL Press, London, pp. 31–52.

Holden, C., Meade, A., Pagel, M., 2005. Comparison of maximum parsimony and Bayesian Bantu language trees. In: Mace, R., Holden, C., Shennan, S. (Eds.), The Evolution of Cultural Diversity: A Phylogenetic Approach. UCL Press, London, pp. 53–66.

Huelsenbeck, J., Rannala, B., 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Systematic Biology 53 (December), 904–913.

Hunley, K., Dunn, M., Lindström, E., Reesink, G., Terrill, A., Norton, H., Scheinfeldt, L., Friedlaender, F., Merriwether, D., Koki, G., Friedlaender, J., 2007. Inferring prehistory from genetic, linguistic, and geographic variation. In: Friedlaender, J. (Ed.), Genes, Language, and Culture History in the Southwest Pacific. Oxford University Press, Oxford, pp. 141–154.

Huson, D., Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution 23, 254–267.

Kirch, P., 1997. The Lapita Peoples. Blackwell, London.

Lindström, E., Terrill, A., Reesink, G., Dunn, M., 2007. The Languages of Island Melanesia. In: Friedlaender, J. (Ed.), Genes, Language, and Culture History in the Southwest Pacific: A Synthesis. Oxford University Press, Oxford, pp. 118–140.

Lynch, J., Ross, M., Crowley, T., 2002. The Oceanic Languages. Curzon Press, Richmond, Surrey.

Matisoff, J., 1990. On megalocomparison. Language 66, 106–120.

McMahon, A., McMahon, R., 2005. Language Classification by Numbers. Oxford University Press, Oxford.

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

Pagel, M., Atkinson, Q., Meade, A., 2007. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. Nature 449, 717–721.

Pagel, M., Meade, A., 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Systematic Biology 53, 571–581.

Paradis, E., Claude, J., Strimmer, K., 2003. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20, 289–290.

Penny, D., Hendy, M., 1985. The use of tree comparison metrics. Systematic Zoology 34, 75–82.

Press, W., 1988. Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press, Cambridge.

Ronquist, F., September 2004. Bayesian inference of character evolution. Trends in Ecology and Evolution 19, 475–481.

Ross, M., 1988. Proto Oceanic and the Austronesian Languages of Western Melanesia. Pacific Linguistics, Canberra.

Ross, M., 2001. Is there an East Papuan Phylum? Evidence from Pronouns. In: Pawley, A., Ross, M., Tryon, D. (Eds.), The Boy from Bundaberg: Studies in Melanesian Linguistics in Honour of Tom Dutton. Pacific Linguistics, Canberra, pp. 301–321.

Spriggs, M., 1997. The Island Melanesians. Blackwell, London.

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

Terrell, J., Kelly, K., Rainbird, P., 2001. Foregone conclusions? In search of ‘‘Papuans’’ and ‘‘Austronesians’’. Current Anthropology 42, 97–124.

Tuffey, C., Steel, M., 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bulletin of Mathematical Biology 59, 581–607.

Wurm, S., 1975. The East Papuan Phylum in general. New Guinea Area Languages and Language Study 1, 783–804.