

## DISCUSSION NOTE

### Typological feature analysis models linguistic geography

MARK DONOHUE

*Australian National University*

SIMON MUSGRAVE

*Monash University*

BRONWEN WHITING

*Australian National University*

SØREN WICHMANN

*Max Plank Institute for Evolutionary  
Anthropology and Leiden University*

Dunn and colleagues (2008) describe and exemplify the use of sophisticated analyses of abstract structural features to reconstruct language histories. The techniques that they use do show some clustering in the groups of languages that they examine; Dunn et al. state that they 'tend to favor a phylogenetic origin for the signal of relatedness' (p. 748), and that the results of their test case 'show a close degree of correspondence to the existing linguistic classification based on sound-meaning correspondences' (p. 747). We argue that a more parsimonious explanation for the results obtained by Dunn et al.'s methodology is that it accurately maps linguistic geography, the network of contact and diffusion that postdates a proto-language, in most cases corresponding to geographic distance.\*

*Keywords:* typology, historical linguistics, computational methods, Austronesian, Papuan

**1. INTRODUCTION: CLAIMS ABOUT TYPOLOGICAL FEATURE ANALYSIS.** The development of new tools that can assist in our understanding of language histories is exciting, and challenging. The prospect of being able to process and empirically evaluate large bodies of data of the sort that has hitherto only been assessed impressionistically is attractive. Nonetheless, new techniques need to be carefully scrutinized before they can be taken up with confidence. Earlier papers (Donohue & Musgrave 2007, Donohue et al. 2008, Donohue 2009) take issue with the interpretation that clusterings based on typological features truly do replicate linguistic phylogenies. In this discussion note we make explicit the problems we see with this approach, and suggest that the analysis of abstract typological features detects (human) geography more robustly than linguistic phylogeny.

Dunn and colleagues (2008; henceforth Dunn et al.) produce a sophisticated cluster analysis based on typological features for two sets of languages, a selection of Oceanic languages from Island Melanesia and a selection of Papuan languages from the same area.<sup>1</sup> Comparing their results to the established phylogeny for the Oceanic languages, they conclude that '[t]he results of the phylogenetic analysis of the sample of Oceanic languages show a close degree of correspondence to the existing linguistic classification based on sound-meaning correspondences' (p. 747).<sup>2</sup> In addition, examining the non-Oceanic languages in their sample with the same methodology, they state that they 'would tend to favor a phylogenetic origin for the signal of relatedness, given the geo-

\* We would like to thank Bernard Comrie, Meladel Mistica, and three referees for *Language* for their input into this publication.

<sup>1</sup> Note that, in contrast to 'Oceanic' (which implies phylogenetic membership in the Oceanic subgroup of the Austronesian family), the term 'Papuan' should not be taken to imply membership in a single family. Rather, the use of the term 'Papuan' merely implies that a language is not Austronesian (or Australian), and that it originates from Melanesia.

<sup>2</sup> Earlier expressed as the claim that the results 'to a very high degree recapitulate the comparative method tree for a branch of the Oceanic languages' (Dunn et al. 2007:389).

graphic distribution of the languages' (p. 748). This leaves us with two questions that can be addressed empirically:

- (i) How close is the replication of the phylogeny of the Oceanic languages?
- (ii) To what extent can geography be ruled out as an explanation for their clustering?

We address both of these questions in the next three sections, and follow that by showing the results of a comparable analysis for a group of related languages of Europe.

**2. REPLICATION.** Dunn et al. claim to show success in replicating the phylogeny presented in Lynch et al. 2002. The tree for the relevant Oceanic languages is shown in Figure 1.

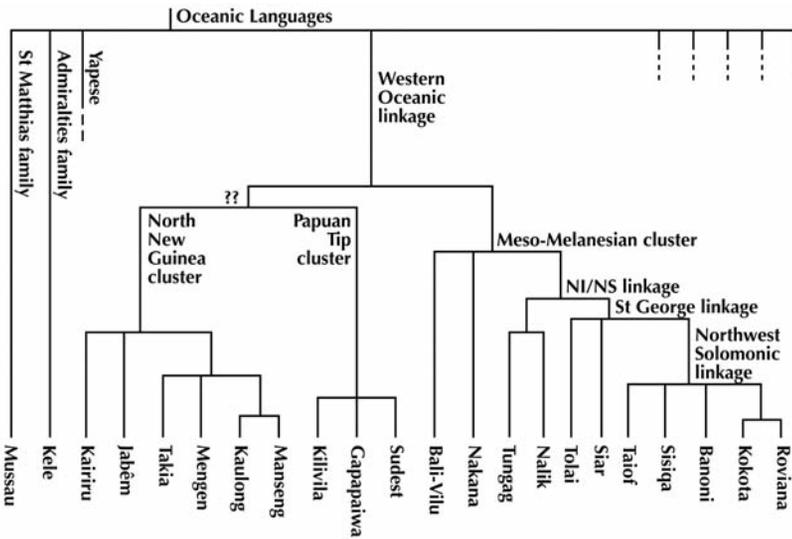


FIGURE 1. Lynch et al. 2002 phylogeny for a selection of Oceanic languages.

It is clear that replication will involve two three-way splits: first, the split into Admiralties, St. Matthias, and Western Oceanic, and second the split of Western Oceanic into North New Guinea, Papuan Tip, and Meso-Melanesian.<sup>3</sup> Other languages of the Oceanic subgroup of Austronesian (e.g. Yapese in Fig. 1) were not included in Dunn et al.'s study, and so should not be considered in any evaluation of replication.

**2.1. EVALUATING REPLICATION.** Dunn et al. present their results in the form of unrooted trees. These trees demonstrate the clustering ('subgroups') of the analysis, without being explicit about a source. We can, nevertheless, quantify a decision about the placement of a root by evaluating which root placement achieves the best replication of the target phylogeny. Placing a root at the position that shows the best replication of the subgroups shown in Fig. 1 makes it easier to evaluate the degree to which hierarchical subgroups of the Oceanic languages are replicated in the clusters formed by application

<sup>3</sup> North New Guinea and Papuan Tip might form a single clade. To complicate matters further, there is evidence that North New Guinea might not itself represent a single clade (the relevant languages are the Schouten group; the question of their status does not affect the present argument). We adopt a conservative position on possible groupings beyond the three-way splits shown in Fig. 1.

of Dunn et al.'s methodology. There are two relevant measures involved, given in 1 and 2, each of which assesses the degree to which subgroups of the data are replicated in the clusters found in the typological analysis (van Rijsbergen 1979).

- (1) Recall: How well does the cluster correspond to (all of) the subset of the data that it is intended to replicate?
- (2) Precision: How well does the cluster exclude the data that does not belong to the subset that it is intended to replicate?

These two metrics can be quantified in terms of true positives (successfully replicated members of a subgroup), false positives (datapoints that should have been excluded from the subgroup, but that were included), and false negatives (datapoints that should have been included in the subgroup, but that were omitted).

$$(1') \text{ Recall} = \text{true positives} / (\text{true positives} + \text{false negatives})$$

$$(2') \text{ Precision} = \text{true positives} / (\text{true positives} + \text{false positives})$$

We demonstrate this heuristic with the data in 3. Here the 1s and 2s represent members of two subgroups, shown in square brackets in 3a. The putative replication in 3b, x and y, can be evaluated to determine the degree to which clusters x and y succeed in replicating the 1 and 2 subgroups. As shown in Table 1, it is trivially true that cluster x completely fails to replicate subgroup 2. By contrast, cluster x shows great precision in replicating subgroup 1, in that it does not include any members of subgroup 2, although only recalling 50% of the members of subgroup 1. Cluster y recalls all of the members of subgroup 2, but also includes two members of subgroup 1, thus showing lower precision. The final column in Table 1 (and subsequently in Tables 2, 3, 4, and 5) shows the evenly weighted harmonic mean ( $F_1$ ) of the recall and precision values for each cluster's replication of each of the two subgroups (assuming equal weight is given to recall and precision). Clearly, and unsurprisingly, cluster x best replicates subgroup 1, and cluster y best replicates subgroup 2; overall, these two best-fit replications (shown without italics in Table 1) score a mean value of 71%.

- (3) a. Subgroups: [ 1 1 1 1 ] [ 2 2 2 ]
- b. Clusters: [x 1 1 ] [y 1 1 2 2 2 ]

		TRUE POSITIVES	FALSE POSITIVES	FALSE NEGATIVES	RECALL	PRECISION	$F_1$
1	x	2	0	2	0.5	1	0.67
	y	2	3	2	0.5	0.4	<i>0.44</i>
2	x	0	2	3	0	0	<i>0</i>
	y	3	2	0	1	0.6	0.75
AVERAGE:							0.71

TABLE 1. Evaluating replication in 3.

**2.2. REPLICATING OCEANIC.** When we apply this methodology to Dunn et al.'s results we find one position in the unrooted tree that results in maximally high replication values for the major subgroups of the Lynch et al. 2002 tree for the Oceanic languages, with an  $F_1$  value of 0.79 (Table 2). Unsurprisingly, this is the point at which the three first-order subgroups are differentiated, shown in Figure 2. (Note that the replication of single nodes, such as Kele or Mussau, is not included in the calculation of overall replication since it is a trivial matter to replicate a single node: every single node in any tree can be replicated, in the sense of containing all and only the members of the single-node subgroup.) While the Meso-Melanesian languages form a single cluster, the Papuan Tip languages Gapapaiwa, Sudest, and Jabêm are split into two clusters; the best replication

comes from the cluster that includes Jabêm as a false positive, and which incorrectly excludes Gapapaiwa. The cluster that best replicates the North New Guinea subgroup is the two-language cluster containing the geographically isolated (in terms of the languages under consideration) Kairiru and Takia (see Fig. 4 for locations of the languages).

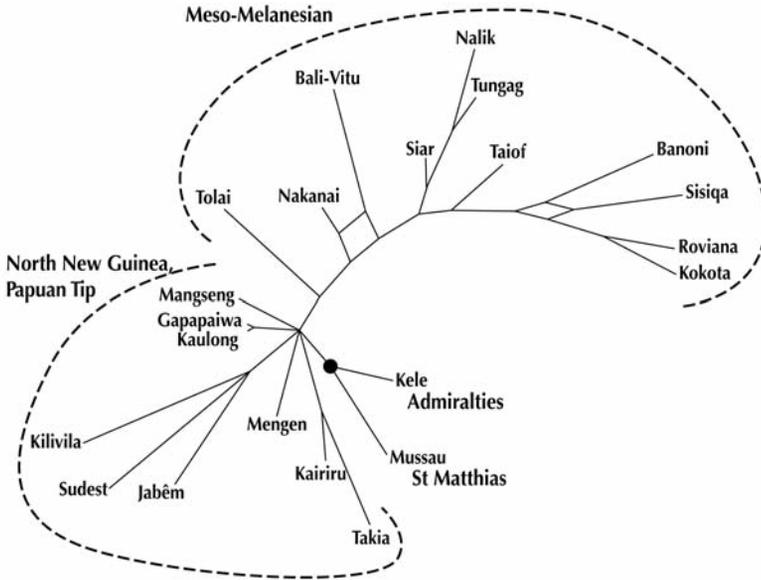


FIGURE 2. The maximally felicitous location for a root in Dunn et al.'s clustering and resulting best-match clusters replicating the groups in Fig. 1.

	TRUE POSITIVES	FALSE POSITIVES	FALSE NEGATIVES	RECALL	PRECISION	F <sub>1</sub>
Western Oceanic	20	0	0	1	1	1
Meso-Melanesian	11	0	0	1	1	1
Papuan Tip	2	1	1	0.67	0.67	0.67
North New Guinea	2	0	4	0.33	1	0.5
				AVERAGE:		0.79

TABLE 2. Replication of the higher levels of the phylogeny in Lynch et al. 2002.

It is clear that, while Dunn et al. achieve a better-than-chance replication of the Lynch et al. 2002 tree, the rate of success is very uneven, as shown in Table 3. The rate of replication ranges from 100% (for Meso-Melanesian, or Western Oceanic as a whole) down to only 50% for North New Guinea. What accounts for these differences?

SUBGROUPS	F <sub>1</sub>
Western Oceanic	0.79
Meso-Melanesian	1.0
Papuan Tip	0.67
North New Guinea	0.5

TABLE 3. Rates of replication for Western Oceanic and its subgroups.

**2.3. A GEOGRAPHIC EXPLANATION.** We note that the different rates of successful replication shown in Table 3 correspond to the increasingly complex phylogeography exemplified by the different Western Oceanic subgroups, shown in Figure 3. Meso-Melanesian is spatially transparent, in the sense that the phylogenetic neighbor of any language is at least one of its geographic neighbors. This is not true for North New Guinea, where (for instance) Kairiru's nearest neighbor is Takia, but these two languages are not especially close phylogenetically, and where Mungen's nearest North New Guinea neighbor, Mangseng, is not particularly close phylogenetically. This strongly suggests that there is not a direct link between the typology-constructed tree and the linguistic phylogeny, but rather that both of them covary according to linguistic geographic distance. Indeed, Dunn et al. note that 'greater structural distance correlates with greater geographic distance' (p. 744).<sup>4</sup>

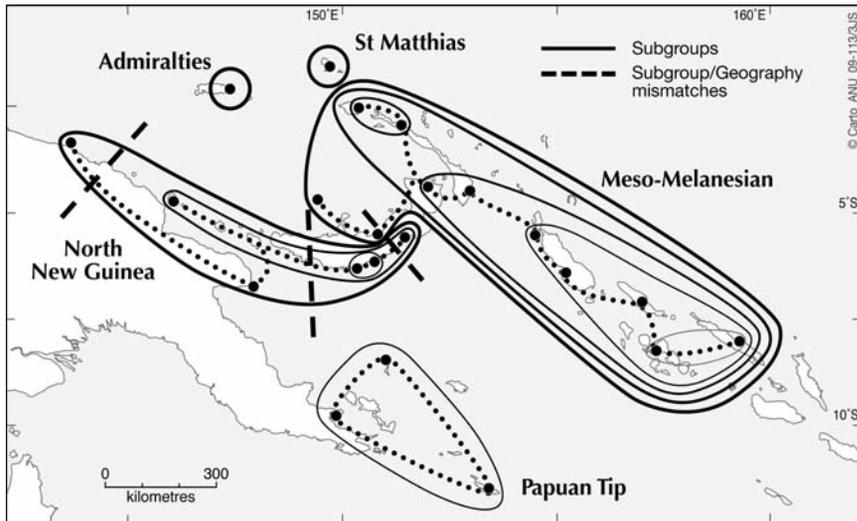


FIGURE 3. Phylogeographies for the Oceanic subgroups.

Following suggestions in Dunn et al. 2007:398, we constructed a network linking the twenty-two Oceanic languages in Dunn et al.'s study, based on social interactions reported in the ethnographic literature. Thus Kele, in the Admiralties, is not linked to Kairiru to its west but to Mussau to the east, mirroring known cultural associations (e.g. Allen 1996, Spriggs 1997)—thus, areas with known cultural relationships are linked, and those with no, or at best tenuous, relationships are not. Table 4 shows the rates of replication that hold between contiguous clusters based on linguistic geography and clusters apparent in Dunn et al.'s study (shown in Figure 4). All of the Dunn et al. clusters represent contiguous blocks on the map in Fig. 4, except for the Gapapaiwa-Kaulong cluster, which is interrupted by Kilivila (as before, singleton groups have not been included because their replication is trivial). Overall, the Dunn et al. clustering replicates the interaction networks illustrated in Fig. 4 at a 96% rate, significantly higher than the 79% rate at which it replicates the Lynch et al. 2002 phylogeny ( $p = 0.10$  comparing the  $F_1$  values for Table 2 and Table 4, two-tailed  $t$ -test;  $p = 0.06$  when comparing recall values alone).

<sup>4</sup> Donohue et al. 2008 first presented the (unquantified) argument that the clustering results presented by an earlier version of Dunn et al. 2008 are best explained by geography.

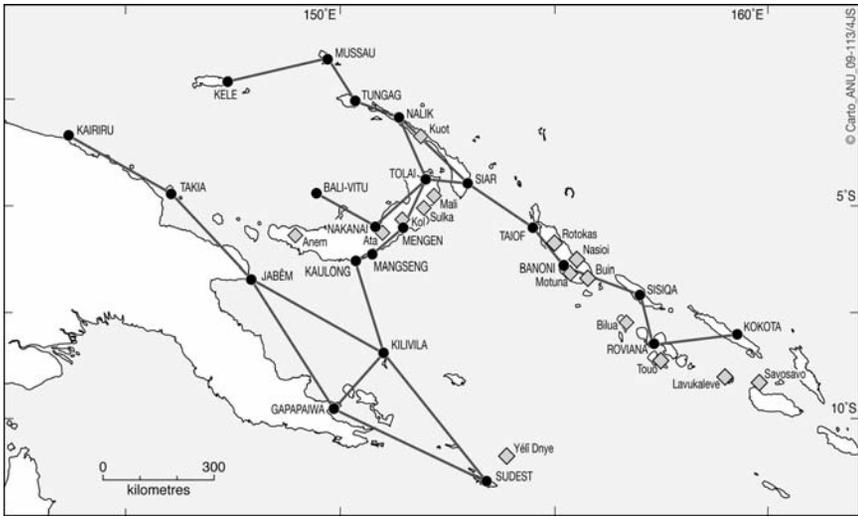


FIGURE 4. An interaction network for the Oceanic languages in Dunn et al. 2008. (Note: Names in all capitals represent the Oceanic languages discussed in this and preceding sections, linked here as a network. Names in normal font show the locations of the Papuan languages discussed in §2.4.)

	TRUE POSITIVES	FALSE POSITIVES	FALSE NEGATIVES	RECALL	PRECISION	F <sub>1</sub>
Kele-Mussau	2	0	0	1	1	1
Kairiru-Takia	2	0	0	1	1	1
Sudest-Kilivila-Jabem	3	0	0	1	1	1
Gapapaiwa-Kaulong	2	1	0	0.67	1	0.8
Meso-Melanesian	11	0	0	1	1	1
AVERAGE:						0.97

TABLE 4. Replication of the nontrivial contiguous geographic clusters in Dunn et al. 2008.

Based on these findings, we believe that the analysis of abstract typological features is a valuable detection tool in that the results serve as an accurate proxy for distance, rather than a proxy for phylogenetic results such as would result from the application of the comparative method to a group of languages.

**2.4. EXPANDING THE TEST CASE: OCEANIC AND PAPUAN COMBINED.** If the claim at the end of the preceding section is true, we would expect to find clusters that represent mixtures of unrelated languages, if geographically intermixed languages from different families were to be combined in the one clustering analysis. This was not considered by Dunn et al., but is not difficult to check given the clarity of the materials they provided for both the Oceanic and Papuan languages in their study. We followed Hunley et al. 2008 and constructed an unrooted tree containing both the Oceanic and Papuan languages in Dunn et al.’s sample. This is shown as the upper panel of Figure 5, produced from SplitsTree (Huson & Bryant 2006) with the data from Dunn et al. 2008. Here the Oceanic languages are shown in ovals, and the Papuan languages in rectangles. It is clear that there is no one clear point that separates all of the Oceanic languages from all of the diverse Papuan languages. This is due to the clustering of Anem, Ata, and Sulka, three Papuan languages of western New Britain, together with the Oceanic languages that come from the west of the study area. Rather than representing a phylogenetic model, the divisions can (and, we argue, should) best be understood in terms of geogra-

phy, as shown in the lower panel of Fig. 5, with languages in the west separate from languages in the east. Within the eastern group we then find a split based on phylogeny, with the Oceanic languages separating from the Papuan languages, and finally within the embedded Papuan cluster we find a secondary geographic split into a northern and a southern group, with Bilua remaining comparatively isolated. Again, Dunn et al. similarly conclude, speaking of the Papuan languages they examined, that ‘the phenetic relationships between the languages (structural distance) clearly correlate with the current geographic position of the languages’ (p. 748).

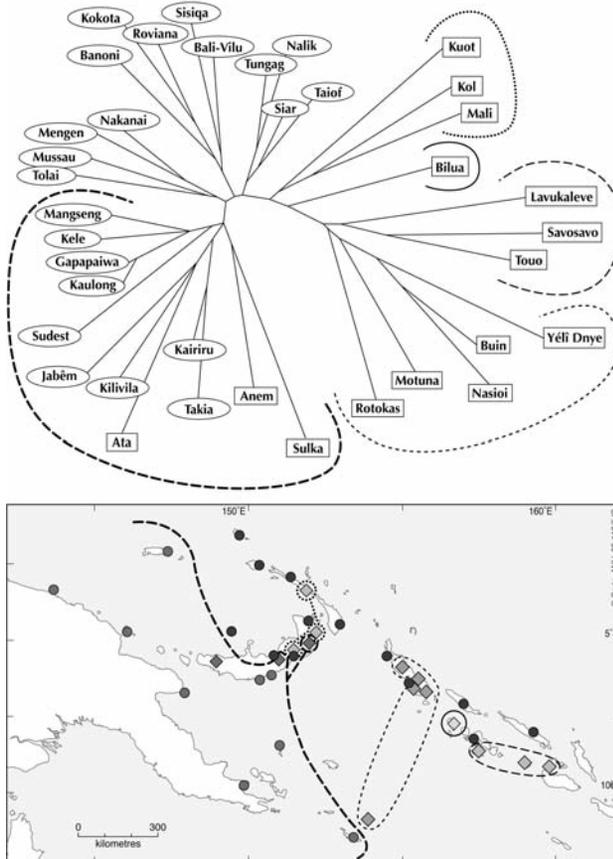


FIGURE 5. Unrooted tree, combined Oceanic and Papuan languages from Dunn et al. 2008 (upper panel), and phylogeography of the clusters (lower panel).

**3. A FAMILIAR EXAMPLE: A ROUGH CALIBRATION BASED ON THE INDO-EUROPEAN LANGUAGES IN EUROPE.** In order to appreciate what a 79% replication rate (see §2) means in terms of achieving a ‘close correspondence’ with an established phylogeny, we tested the results of a typological clustering against an accepted phylogeny. We selected those Indo-European languages of Europe that are best attested in the *World atlas of language structures* (WALS; Haspelmath et al. 2005) and that are all found in Europe, in order to replicate the geographical compactness and phylogenetic unity of Dunn et al.’s test case with Oceanic languages of Island Melanesia (following Nichols & Warnow 2008:777).<sup>5</sup>

<sup>5</sup> Only about 40% of the features in WALS are shared with Dunn et al.’s selection of features, which were selected both to provide ‘broad typological coverage’ and because the particular features are ‘expected to dis-

After sorting the information in *WALS* to exclude lexical comparisons, to match the methodology found in Dunn et al. 2008, we arrived at a test sample of twenty-one languages for which between fifty and 128 features were coded per language (the data is presented in the appendix). The resulting character matrix was then clustered using SplitsTree, with the result as shown in the neighbor-joining tree in Figure 6; numbers show values for 10,000 bootstrap replications. Figure 7 shows a NeighborNet representation for the same data.<sup>6</sup> The main discrepancies that this tree shows with a generally accepted representation of Indo-European relationships, highlighted in Fig. 7, are:

- the loss of Icelandic from North Germanic,
- the inclusion of English within North Germanic (and exclusion from West Germanic),
- the inclusion of Romanian with Albanian (and exclusion from Romance), and
- the inclusion of Bulgarian in a clade with Greek, and its loss from Slavic.

Table 5 quantifies the degree of match; at 85%, the replication rate (under the same condition of discounting trivial singleton replications) is higher than the 79% achieved by Dunn et al. for the Oceanic languages in their sample. If we were to group, for instance, Polish with Dutch and German, the resulting topology would match the 79% replication rate that was found in Table 2 for the replication rates in Dunn et al. 2008. This allows the reader to judge the fidelity shown by a 79% replication rate.

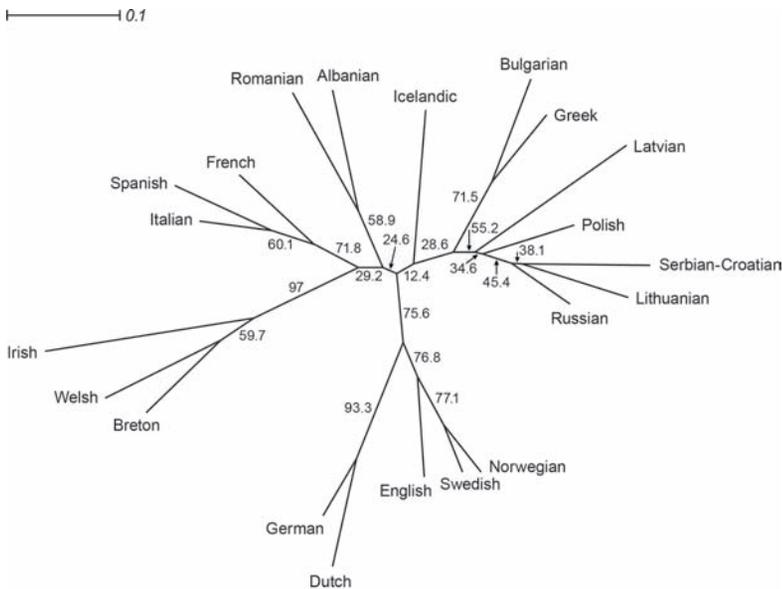


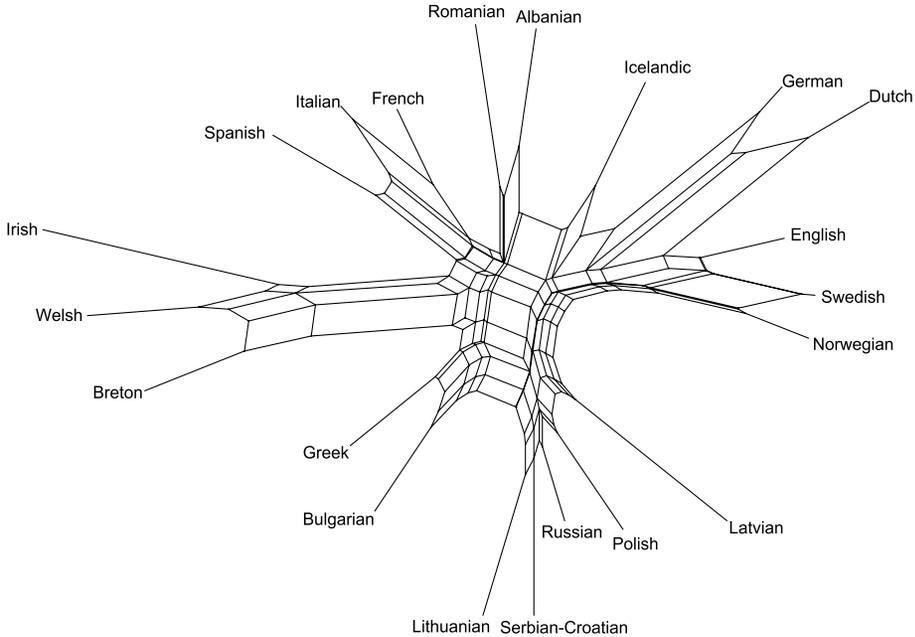
FIGURE 6. Typological clustering of twenty-one languages of Europe (data from *WALS*).

tinguish between the languages of Island Melanesia', specifically features 'in phonology, morphology, and syntax on which Austronesian and Papuan languages generally diverge' (p. 730). (Note that these features fail to separate the Oceanic (= Austronesian) and Papuan languages, as we have shown in Fig. 5.) We believe that the *WALS* features similarly provide 'broad typological coverage', and differ from the Dunn et al. selection by not being biased toward a particular result.

<sup>6</sup> Dunn et al. do not use SplitsTree, but a Bayesian algorithm. Leaving aside questions about the strength of conclusions based on Bayesian techniques (e.g. Huelsenbeck et al. 2001, Yang 2008), we note that our clustering based on SplitsTree shows a 91% match with that of Dunn et al. in the relative topology of the Oceanic languages. Further, an extremely similar clustering topology, including the mixed Papuan-Oceanic cluster highlighted in our Fig. 5, is reported in Hunley et al. 2008.

	TRUE POSITIVES	FALSE POSITIVES	FALSE NEGATIVES	RECALL	PRECISION	F <sub>1</sub>
Germanic	5	0	1	0.83	1	0.91
Romance	3	0	1	0.75	1	0.86
Albanian	1	0	0			
Balto-Slavic	5	0	1	0.83	1	0.91
Slavic	3	1	1	0.75	0.75	0.75
Celtic	3	0	0	1	1	1
Greek	1	0	0			
AVERAGE:						0.85

TABLE 5. Quantifying replication of the phylogeny of the Indo-European languages in Fig. 7.

FIGURE 7. NeighborNet clustering of twenty-one languages of Europe (data from *WALS*).

In all cases the points of divergence from the received view on Indo-European relationships have trivial explanations.

- Icelandic is geographically maximally distant from the other European languages, and has not shared in many of the changes that have characterized its nearest cousins.
- English is included more closely with the North Germanic languages than with its continental cousins because it has undergone extensive Scandinavian influence, and is separated from the mainland Germanic languages by the English Channel.
- Romanian is at a great remove from the rest of the Romance family, and has been extensively influenced by the surrounding languages.
- Greek and Bulgarian are neighbors sharing a long history of interaction.

In every case for which phylogeny does not straightforwardly correspond to geography, we find a difference between the subgroups in the traditional linguistic classification and the clusters arrived at by analysis of typological features. In short, linguistic geography, rather than phylogenetic identity, determines typological clusters. If linguistic geography corresponds to physical geography, as is common, then a clustering that replicates geography will give the illusion of the replication of a phylogeny.

Additionally, if the tree in Fig. 6 is taken to be an accurate replication of the phylogeny of Indo-European languages, then we can similarly have confidence that the method described by Dunn et al. replicates phylogenies. If we hesitate to accept the tree in Fig. 6 as an accurate portrayal of the Indo-European languages of Europe, we should similarly hesitate to accept the claims about the phylogenetic value of typological analysis unconstrained by formal correspondences.

**4. CONCLUSIONS.** Unlike Dunn et al., we do not believe we ‘can be agnostic about inheritance vs. diffusion’ when interpreting the results of a clustering analysis based on typological features. While their results show a clearly better-than-chance replication of the groups and subgroups of Oceanic languages in their sample, we have argued that this is not a direct correspondence, but is achieved through the common factor of linguistic geography.<sup>7</sup> Linguistic phylogenies typically disperse through contiguous regions, and typically remain contiguous following dispersal. For this reason a model that claims to replicate a linguistic phylogeny can ONLY be shown not to detect linguistic geography, the spatially measured network of social interactions, by examining the accuracy of both replication of noncontiguous clades in a phylogeny and of spatially isolated clades that are geographically associated with another, phylogenetically distant, clade. When we examined the discrepancies between Dunn et al.’s results and the linguistic phylogenies against which they were evaluated, we found that in every case either a geographically noncontiguous clade was not successfully clustered together, such as the failure to cluster Takia closely with the North New Guinea languages of New Britain, or a spatially isolated clade was clustered with one of its neighbors, such as the cluster that contains both Kele and Mussau (compare Fig. 1 with Fig. 2). This clearly indicates that, if there is a mismatch between subgroups formed by inheritance (the phylogeny) and those formed by the diffusion (spatial or social) of linguistic traits, the methodology unambiguously detects diffusion, as would be expected given the universal tendency for abstract typological features to follow spatial distributions (Holman et al. 2007, though see Nichols 2003 on areally vs. genetically stable typological features).

We do not argue with the result that shows that different linguistic families show different typological profiles, and we do not suggest that this approach has nothing to say in the detection of language families (cf. Wichmann & Saunders 2007, Donohue & Whiting 2009). We do argue that using typological analysis in an attempt to detect phylogenetic signals (Dunn et al. 2008:745), in the absence of formal similarities in the phonological shape of morphemes and lexical items, cannot be assumed to be revealing of patterns of genealogical relationship that are not more clearly detected by application of the comparative method or even by lexicostatistics, unless geographical factors can be eliminated as a simpler explanation. Moreover, in any such undertaking it is a minimal requirement to test the performance of the method on a well-known family, such as Indo-European; our preliminary examination of this well-understood family is not encouraging.<sup>8</sup> Figure 8 evaluates the different nodes in Fig. 6 in terms of their success at replicating a clade of Indo-European—fully replicating, meaning both (i) including all

<sup>7</sup> We note in passing that Baayen (2009:142–48) reaches a similar conclusion: ‘the unrooted phylogenetic tree groups languages according to geographic region’ (p. 145).

<sup>8</sup> Similarly, we are unclear why a time period ‘older than 3,200 years’ is ‘a period that is otherwise inaccessible to historical linguistics’ (Dunn et al. 2008:748). We understand that, referring to the Papuan languages of Island Melanesia, ‘A stumbling block to any claim of relatedness of these languages is the fact that, apart from some small lower-level families, they lack the lexical and morphological correspondences that would allow an unequivocal identification of cognates, affording the extraction of systematic sound changes and thus eventual reconstruction of ancestral languages’ (p. 747), but suggest a simpler explanation: the Papuan languages represent a mosaic of mutually unrelated language families (following Ross 2001).

members of the clade and (ii) excluding all nonmembers—compared to the bootstrap values associated with each node. There are only three fully successful replications: Brythonic, replicated with a bootstrap value of 59.7; Western Romance, 71.8; and Celtic, 97 (note that Celtic, the only successfully replicated first-order subgroup of Indo-European in this study, is also the most geographically compact group in the sample). Figure 8 groups the bootstrap values in bins of twenty, so the first value represents the ratio of the number of nodes that successfully replicate a clade to those that do not, for bootstrap values between 0 and 20. In this case there is only one member of the bin, and it does not replicate any clades. The same is true for the 21–40 bin, but in each of the remaining bins we find one successful replication, representing one third, one fifth, and then half of the nodes with that level of bootstrap value (moving rightward along the figure) successfully replicating established subgroups of languages. It is clear from this figure that, to have a 50% chance of successfully replicating a known group or subgroup in a family, the bootstrap values must be above 80 (our data has no scores between 81 and 90, so we cannot judge more finely).<sup>9</sup> Bootstrap values above ~50 can be seen, from Fig. 7, to reliably replicate geographically contiguous regions. The only geographically discontinuous subgroup in the sample, Romance (with French, Italian, and Spanish separated from Romanian by Slavic), is not replicated except in a cluster that includes Albanian and Celtic.

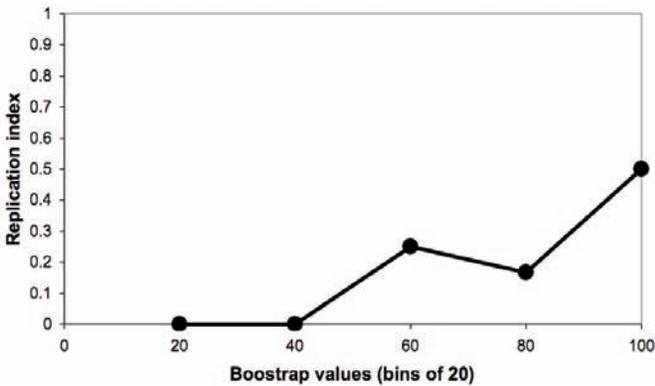


FIGURE 8. Successful replication of Indo-European clades in Dunn et al.'s figure 7, by bootstrap values.

Applying this heuristic to Dunn et al.'s tree in their figure 7 (our Fig. 2), we find very few nodes with more than two languages, geographically contiguous, grouped with such bootstrap values (Table 6).

	BOOTSTRAP VALUE	PHYLOGENETIC UNIT?	GEOGRAPHICALLY CONTIGUOUS?
Roviana–Kokota	85	yes	yes
Banoni–Sisiqa–Roviana–Kokota	80	no	yes
Nalik–Tungag	96	yes	yes
Kele–Mussau	100	no	yes
Gapapaiwa–Kaulong	77	no	no
Nakanai–Bali–Vitu–Siar–Nalik–Tungag– Taiof–Banoni–Sisiqa–Roviana–Kokota	83	no	no

TABLE 6. Evaluating nodes with high (> 75) bootstrap values from Dunn et al.'s figure 7.

<sup>9</sup> We are not clear on the process of reweighting that Dunn et al. apply to their bootstrap values, and so cannot comment on this process and the results it produces.

The results shown in Fig. 7 suggest that replication values below approximately 0.8 ~ 0.9 do NOT count as 'successful replication', that nodes with bootstrap values below 0.75 are not convincing, and that even those nodes with high bootstrap values are, with the kind of character set employed by Dunn et al., more likely to replicate geographic regions than genealogical clades (echoing Anttila 1989:321).

## APPENDIX

Below we present the nexus file used to generate the clustering analysis shown as Figs. 6 and 7, run through SplitsTree (Huson & Bryant 2006). The characters of data are included in the 'Matrix' section, and comprise the first 128 features coded in *WALS* (Haspelmath et al. 2005), including all phonological, morphological, and syntactic features (excluding only the lexical features). A description of the features, in the same order as listed here, is available at <http://wals.info/feature>.

```
#nexus

BEGIN Taxa;
DIMENSIONS ntax=21;
TAXLABELS
[1] 'Serbian-Croatian'
[2] 'Norwegian'
[3] 'Breton'
[4] 'Welsh'
[5] 'Icelandic'
[6] 'Swedish'
[7] 'Italian'
[8] 'Dutch'
[9] 'Lithuanian'
[10] 'Bulgarian'
[11] 'Romanian'
[12] 'Albanian'
[13] 'Polish'
[14] 'Irish'
[15] 'German'
[16] 'Latvian'
[17] 'Russian'
[18] 'French'
[19] 'Greek'
[20] 'Spanish'
[21] 'English'
;
END; [Taxa]

BEGIN Characters;
DIMENSIONS nchar=128;
FORMAT
  datatype=STANDARD
  missing=?
  gap=-
  symbols="0 1 2 3 4 5 6 7 8 9"
  labels=left
  transpose=no
  interleave=yes
;
MATRIX
'Serbian-Croatian' ?????????????1555????????????2?????26??55????1??1??561?7???4???
'Norwegian'      33222112??232144411?????2?????2??332?????2?1??117???4???
'Breton'         43342112??231681?11?????3?????2??11?????2?????717???2???
```

'Welsh' ??????????681????????23?????14??22?32?1311937??????  
'Icelandic' ??????????2811?????222??32326?334??????12?46137??4??  
'Swedish' ??????????1445????????2??2??26??31??21??231?23917??4??  
'Italian' ??????????1364????????23?????26?412??21?3222213917??4??  
'Dutch' ??????????1431????????2??2??26?411?????32312139?61?24??  
'Lithuanian' 52442113??13?156411?????2??2??26??55??21??312?631375?????  
'Bulgarian' 52342112??131????11?????2?????2??21??????12?139?751?4??  
'Romanian' 33342112??1?1136?11?????2?????26??32?????312?231165?4??  
'Albanian' 333421123?331??15?????2?????2??431?????21??24611651?4??6  
'Polish' 42542112??131681111?????2?????2??155?????22122661275??4??  
'Irish' ???????32??2824????????4332??664414333133121322517??4??  
'German' 331423122223114311112222233232326431233115324224611651?421?  
'Latvian' 2234211232132281511?????3?32223264?223321522122561175114??  
'Russian' 425421123213115651112322233323264553321?321226613751?4216  
'French' 331423123123113541115342523322232664123312532222119171114216  
'Greek' 3234211232131586115123333233232654123331?32122361175??421?  
'Spanish' 423311123212113711511332523322232654123331112322139?71?421?  
'English' 33142112221311441151522223223222644113221231212239171124216  
'Serbian-Croatian' ??1?????5114?211??31212?23111?15?14?3?????????2??2?115221  
'Norwegian' ???????24142??????212?2311111611443??1?????????2??4??  
'Breton' ??1?????21?????11??322?222411?111444??2?????????1??11221  
'Welsh' ?3?????2124?232????32212222113111444??2?221??1?????2??11211  
'Icelandic' ?51?212122142212????212?2211111??443??1??????38??173??221  
'Swedish' ?51?212124142211261121212111111611443??1?????????2??345221  
'Italian' ???1?????2234?211????232?22211116?1444??222211?1?????21736??  
'Dutch' ?31?????2414?2112511713622111?6?1555??21221??1?????2??245221  
'Lithuanian' ??1?????2114?21122??212?211111111443??2??3?????1??11?222  
'Bulgarian' ?31?112321412??331173212311111311443??2????3?????2??11?221  
'Romanian' ??1?11242231?21111??212?2226111611444?????????????2??165221  
'Albanian' 531?????2221121222??212?22211111?1444??22523??1?????217?15221  
'Polish' ??1?????2114?211??31232?2211111111443??23221?31?????2??115222  
'Irish' 561?????4121221111313221222211?1114441226221??1??21233111211  
'German' ?3?12121221422112511713622111116115552221221131381221724??  
'Latvian' ?4112113431422112211232?211111111144322212411?13822133111222  
'Russian' 23111124411422111131212?221111131144322212211313822?17111222  
'French' 23111112142211253121212221111111444122122113138?421731?221  
'Greek' ?3111121214122111131732?22111111114432225231313821217115221  
'Spanish' 23111111213422111131232122211116114442225231313834217345211  
'English' 66112121251422111112121231111161144312212214213812233345221

'Serbian-Croatian' ???????  
'Norwegian' ???????  
'Breton' 1??????  
'Welsh' ???????  
'Icelandic' ???????  
'Swedish' ???????  
'Italian' ?11?322?  
'Dutch' 4??????  
'Lithuanian' ???????  
'Bulgarian' ?112????  
'Romanian' ???2????  
'Albanian' 4??2????  
'Polish' ???????  
'Irish' 443?2221  
'German' ?111221?  
'Latvian' 4111????  
'Russian' 41112221  
'French' 41113222  
'Greek' 41123111

```

'Spanish'          44113222
'English'         41112221
;
END; [Characters]

BEGIN st_Assumptions;
  uptodate;
  chartransform=Uncorrected_P HandleAmbiguousStates = Ignore;
  disttransform=NeighborNet Variance = OrdinaryLeastSquares Threshold = 1.0E-6;
  splitstransform=EqualAngle;
  SplitsPostProcess filter=dimension value=4;
  autolayoutnodelabels;
END; [st_Assumptions]

```

## REFERENCES

- ALLEN, JIM. 1996. The Pre-Austronesian settlement of Island Melanesia: Implications for Lapita archaeology. *Prehistoric settlement of the Pacific*, ed. by Ward H. Goodenough, 11–27. Philadelphia: American Philosophical Society.
- ANTTILA, RAIMO. 1989. *Historical and comparative linguistics*. Amsterdam: John Benjamins.
- BAAYEN, HARALD. 2009. *Analyzing linguistic data*. Cambridge: Cambridge University Press.
- DONOHUE, MARK. 2009. Geography is more robust than linguistics. *Science* e-letter, 13 August 2009. Online: <http://www.sciencemag.org/cgi/eletters/324/5926/464-c>.
- DONOHUE, MARK, and SIMON MUSGRAVE. 2007. Typology and the linguistic macro-history of Island Melanesia. *Oceanic Linguistics* 46.2.348–87.
- DONOHUE, MARK, and BRONWEN WHITING. 2009. Assessing California as a linguistic area. Paper presented at the Workshop on American Indigenous Languages, University of California, Santa Barbara, 8–9 May 2009.
- DONOHUE, MARK; SØREN WICHMANN; and MIHAI ALBU. 2008. Typology, areality and diffusion. *Oceanic Linguistics* 47.1.223–32.
- DUNN, MICHAEL; ROBERT FOLEY; STEPHEN C. LEVINSON; GER REESINK; and ANGELA TERRILL. 2007. Statistical reasoning in the evaluation of typological diversity in Island Melanesia. *Oceanic Linguistics* 46.2.388–403.
- DUNN, MICHAEL; STEPHEN C. LEVINSON; EVA LINDSTRÖM; GER REESINK; and ANGELA TERRILL. 2008. Structural phylogeny in historical linguistics: Methodological explorations applied in Island Melanesia. *Language* 84.4.710–59.
- HASPELMATH, MARTIN; MATTHEW S. DRYER; DAVID GIL; and BERNARD COMRIE. 2005. *World atlas of language structures*. Oxford: Oxford University Press.
- HOLMAN, ERIC W.; CHRISTIAN SCHULZE; DIETRICH STAUFFER; and SØREN WICHMANN. 2007. On the relation between structural diversity and geographical distance among languages: Observations and computer simulations. *Linguistic Typology* 11.2.395–423.
- HUELSENBECK, JOHN P.; FREDERIK RONQUIST; RASMUS NIELSEN; and JONATHAN P. BOLLECK. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294.2310–14.
- HUNLEY, KEITH; MICHAEL DUNN; EVA LINDSTRÖM; GER REESINK; ANGELA TERRILL; MEGHAN E. HEALY; GEORGE KOKI; FRANÇOISE R. FRIEDLAENDER; and JONATHAN S. FRIEDLAENDER. 2008. Genetic and linguistic coevolution in Northern Island Melanesia. *PLoS Genetics* 4.10.e1000239. Online: <http://www.plosgenetics.org/article/info:doi/10.1371/journal.pgen.1000239>.
- HUSON, DANIEL H., and DAVID BRYANT. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23.2.254–67. Software available from <http://www.splitstree.org>.
- LYNCH, JOHN; MALCOLM ROSS; and TERRY CROWLEY. 2002. *The Oceanic languages*. London: Curzon.
- NICHOLS, JOHANNA. 2003. Diversity and stability in language. *The handbook of historical linguistics*, ed. by Brian D. Joseph and Richard D. Janda, 283–310. Oxford: Blackwell.
- NICHOLS, JOHANNA, and TANDY WARNOW. 2008. Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass* 2.5.760–820.

- ROSS, MALCOLM. 2001. Is there an East Papuan phylum? Evidence from pronouns. *The boy from Bundaberg: Studies in Melanesian linguistics in honour of Tom Dutton*, ed. by Andrew Pawley, Malcolm Ross, and Darrel Tryon, 301–21. Canberra: Pacific Linguistics.
- SPRIGGS, MATTHEW. 1997. *The Island Melanesians*. Oxford: Blackwell.
- VAN RIJSBERGEN, C. J. KEITH. 1979. *Information retrieval*. London: Butterworth.
- WICHMANN, SØREN, and ARPIAR SAUNDERS. 2007. How to use typological databases in historical linguistic research. *Diachronica* 24.2.373–404.
- YANG, ZIHENG. 2008. Empirical evaluation of a prior for Bayesian phylogenetic inference. *Philosophical Transactions of the Royal Society* B363.4031–39.

[mark.donohue@anu.edu.au]

[Simon.Musgrave@arts.monash.edu.au]

[bronwen.whiting@anu.edu.au]

[wichmann@eva.mpg.de]

[Received 3 November 2009;

accepted pending revisions 17 May 2010;

revision received 8 July 2010;

accepted 23 November 2010]