Squib

Typology, Areality, and Diffusion

Mark Donohue, Søren Wichmann, and Mihai Albu

AUSTRALIAN NATIONAL UNIVERSITY, MAX PLANCK INSTITUTE FOR EVOLUTIONARY ANTHROPOLOGY/LEIDEN UNIVERSITY, AND MCMASTER UNIVERSITY

Dunn et al. (2007) state that their typological comparisons do not demonstrate genealogical relatedness in the usual sense, but that the technique does accurately recapitulate trees established by the comparative method. We demonstrate that the signal picked up by their method is areal, rather than genealogical, and suggest that the method, when tested on known language families, will also show a high sensitivity to the effect of diffusion.

1. THE APPLICATION OF TYPOLOGY IN HISTORICAL LINGUISTICS.[1] In their response to Donohue and Musgrave (2007), Dunn et al. (2007) defend the stance that typological information about languages should be considered a tool that can be used to measure the phylogenetic relatedness of languages. Despite introducing a distinction between linguistic relatedness, as it is commonly understood, and historical relationships (which include contact-induced similarities), they maintain the position that typological features can identify which languages share common ancestors and which do not. We raise doubts about the claim that trees drawn on the basis of typological similarities reveal the same kind of information that is generated by the application of the comparative method, and suggest that Dunn et al. have captured an areal signal. This means that there is as yet no data to support the claim of plausibility for the hypothesis that the 'East Papuan' languages represent a genealogical unit (see Ross 2001a).

2. WHAT DUNN ET AL. CLAIM. Dunn et al. (2007) clarify their position on the thrust of Dunn et al. (2005) as being that "the computational reconstruction of language history using typological features offers a new and exciting prospect for understanding language prehistory" (2007:389) (a point with which we agree), and that "we do not claim to have shown that the Island Melanesian Papuan languages have one linguistic ancestor" (2007:390). This second claim is technically accurate; the 2005 article contains quotes such as "[a] plausible interpretation of the Papuan language tree is that the two language groups now located on the Solomons and Bougainville separated from a common ancestor" (2005:2075), and "[t]he most plausible hypothesis to explain this result is the

1. Thanks to Juliette Blevins, Bernard Comrie, Russell Gray, Johanna Nichols, and Vladimir Polyakov for suggestions that improved the paper, and to Thomas Mailund for computational help with tree comparisons.

divergence of the Papuan languages from a common ancestral stock" (2005:2072). Technically there is no claim in the 2005 article, but there is a strong suggestion.

Dunn et al. (2007:388) later state that "the phylogenetic relatedness of the Papuan languages remains a serious hypothesis", and that "Dunn et al. (2005) presented the first evidence that there can be phylogenetic information in linguistic typological data" (2007:395). It is unclear what the claim is (or was). More importantly, do they demonstrate this claim? Donohue and Musgrave (2007) raised doubts, which Dunn et al. (2007) largely did not address. In this article we elaborate on some doubts concerning the applicability of this technique to the establishment of genealogical relationships.

3. LINGUISTIC FEATURES. The sample of 125 features in Dunn et al. (2005) covers a wide range from phonology to discourse. The authors state that "[l]ogically dependent features were eliminated", but do not mention how they decided what the logically dependent features were—by fiat or by some consistent method. (For a non-impressionistic method, see Holman 2008.)

If traits were selected on the basis of the known typology of the region (Dunn et al. 2007:391), the results of the comparison cannot be trusted unless they have been independently validated against an external group. It is clear from numerous studies that one can skew data in whatever direction one wants to by selecting the 'right' features. To give a short example from more familiar Western European languages, on the basis of typological features Dutch can be grouped with either English, to which it is known to be closely related, or French, with which it shares a close proximity, depending on the features selected for comparison. Table 1 shows that while there are some features that Dutch shares with English to the exclusion of nearby Romance languages, and some features that French shares with other Romance languages to the exclusion of Dutch, it would not be hard to select features that group Dutch with French (for instance, by selecting features 4, 5, 6, and 7). (Convincingly grouping Dutch with Spanish would not be so trivial, though feature 10 is a start.) The number of features selected can be expanded until the illusion of an independent, representative sample is achieved.

TABLE 1. DIFFERENT FEATURES RESULT IN DIFFERENT GROUPINGS IN WESTERN EUROPE

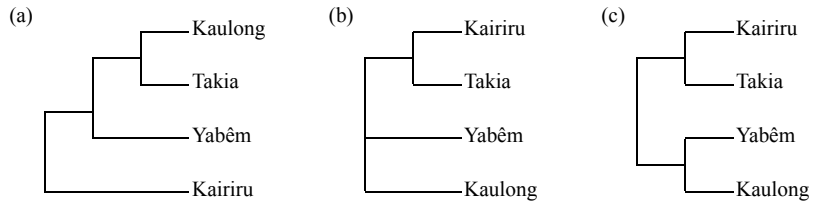
	ENGLISH	DUTCH	FRENCH	SPANISH
1. Case on independent 3rd person pronouns	√	√	?	
2. Language has phonemic *h*	√	√		
3. AdjN order	√	√		
4. Uvular rhotics		√	√	
5. Language has phonemic *y*, *ø*, or *œ*		√	√	
6. Agreement on adjectives		√	√	√
7. 2SG politeness contrasts		√	√	√
8. NAdj order			√	√
9. Agreement for object			√	√
10. Language has phonemic *x* or *γ*		√	——	√
11. Language has phonemic *θ*	√			√

4. CALIBRATION AND AREALITY. In order to demonstrate the validity of a new technique for determining linguistic relations, we need to first show that it can successfully reproduce the results that are obtained by the comparative method. Ideally this should involve an area in which linguistic relations are beyond question: examining, for instance, the internal relations of the Romance languages. Dunn et al. state that "we found that we could to a very high degree recapitulate the comparative method tree for a branch of the Oceanic languages" (2007: 389; see also p. 401), while acknowledging that there is still controversy in the subgrouping of the Western Oceanic languages that they examine. For instance, comparing Ross (1988) with Lynch, Ross, and Crowley (2002) reveals substantial changes in the higher-order groupings, and in the internal organization of Papuan Tip languages. The Western Oceanic languages remain the least studied of the Oceanic languages, and have a history of significant contact-induced change effected by the neighboring Papuan languages (Ross 2001b), which makes it somewhat unfortunate as a proving ground for the calibration of a new methodology.

Examining the recapitulation, there are a number of problems. Within the Meso-Melanesian cluster, Dunn et al. (2005) posit a number of subgroups that are not part of Ross's classification. Subgrouping within the Papuan Tip cluster is uncertain, and the flat structure represents expert opinion as well as any structure. The representation of the North New Guinea languages, however, is anomalous. Dunn et al.'s (2005) typological methodology results in an unrooted tree in which Kairiru is sister to Takia, and Yabêm is sister to Kaulong, as shown in (c) in figure 1. This is presented as "showing a high degree of concordance" with Ross's classification, represented as having Kairiru the sister of Takia, and the other two languages joining at a higher level (2005:2074). This description of Ross's classification in Dunn et al. (2005) is shown as (b) in figure 1; this classification matches figure 1 in Dunn et al. (2007). The classification actually described for the North New Guinea languages in Ross (1988) is shown as (a) in figure 1. Here the low-level sister of Takia is *not* Kairiru, but rather Kaulong. Taking into account Ross's representation of the North New Guinea linkage, this represents a serious difference between the typological tree and the comparative-method tree (see later in this section for a quantified account of the "degree of concordance" between the two trees).

What accounts for this difference in tree topology? The answer is, simply, geography. The nearest neighbor to Kairiru is Takia, and Dunn et al.'s methodology has linked the two

FIGURE 1. THREE CLASSIFICATIONS OF FOUR NORTH NEW GUINEA CLUSTER LANGUAGES †

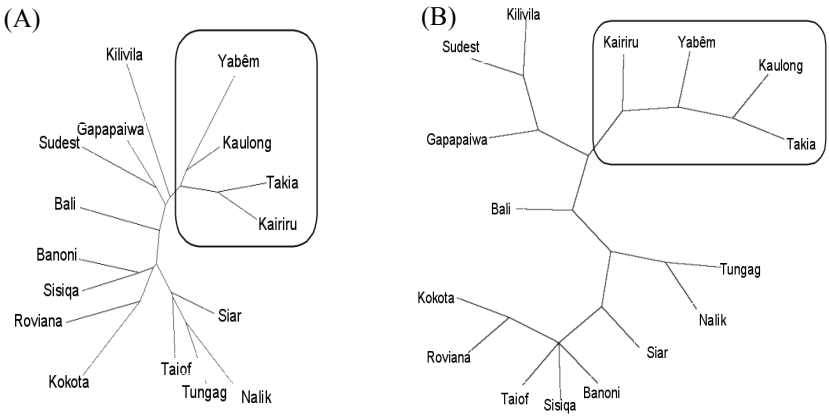


† (a) Ross (1988); (b) Dunn et al.'s (2005) citation of Ross (1988), and Dunn et al. 2007; (c) Dunn et al. (2005)

languages on the basis of proximity. The same principle, linking languages that are close to each other, also accounts for the unmotivated linking of Yabêm and Kaulong, and (in the Meso-Melanesian subgroup) of Banoni with Sisiqa (see figure 2). Clearly (c) in figure 1 matches (b) much better than it matches (a); this error of reporting raises doubts about how closely Dunn et al.'s tree does recapitulate the classification derived from the comparative method. Figure 2 shows the Western Oceanic languages used by Dunn et al. in an unrooted tree produced from their typological features together with an unrooted tree from Ross (1988). While the differences in branch lengths are not significant, since these have been standardized in the Ross tree, and while the position of the Kokota-Roviana group to the left or the right of Sisiqa is also not significant, it is clear that there are a number of points of discrepancy in the topologies of the two trees. How great are these differences?

In order to measure the magnitude of the differences we applied a simulation strategy, as follows. By means of a computer program we generated 10,000 abstract phylogenetic trees, all having sixteen taxa. In order to approach a realistic set of trees we employed a branching model for phylogenies, described by Chu and Adami (1999), that builds on a standard model developed in the nineteenth century (known as the Galton-Watson process). Briefly, the model operates with different probabilities for different numbers of offspring of a given taxon: a certain probability for zero offspring, a certain probability for one offspring, another for two offspring, and so forth. In the situation where the probabilities add up such that a growth is certain, the process is known to result in a distribution of phyla that resembles that of the world's language families: a few very big ones, some intermediate ones, and a lot of small ones, more precisely a so-called power law distribution (Wichmann 2005). Thus the random trees could be viewed as 10,000 ways in which a family of sixteen members could have developed from one and the same ancestor

FIGURE 2. WESTERN OCEANIC LANGUAGES REPRESENTED IN UNROOTED TREES AS REPORTED (A) BY DUNN ET AL. (2005), AND (B) ROSS (1988) †



† The North New Guinea cluster languages have been marked with a box; note also the different topology for the Meso-Melanesian languages at the opposite end of the tree (e.g., Taiof)

according to a realistic branching process. All pairs among the 10,000 trees were then compared and the Robinson-Foulds (RF) distances were calculated. The RF distance is a simple count of the number of nodes that are found in one tree but not the other (comparing first one tree with the other, then the other way around, and finally dividing by two).

The RF distance between the tree produced by Dunn et al. (2005) and Ross's tree as represented by them is 4; a distance of 4.5 will occur with a probability of 0.05 between any two trees in the random sample, and so it can be said that the match between the two trees is much greater than would be expected by chance. The Ross tree as represented by Dunn et al. is not, however, an accurate representation of Ross (1988). Comparing Dunn et al.'s tree with the correct tree, the RF distance is 7. An RF distance of 7 corresponds to the median of distances among random trees; this is exactly the degree of similarity that would be expected by chance between any two trees with the same sixteen taxa. In Dunn et al. (2007) some improvements in the topology of the tree were made by applying a Bayesian algorithm rather than a parsimony-based one, and this brings the RF distance down again, to 4.5. The trees are now more similar than would be expected by chance, but a better match should be expected for a method that is supposed to handle cases such as the Papuan one where the time depth would be many times more than the time depth of the Oceanic subgroup of Austronesian. While the RF method is somewhat simplistic inasmuch as it exaggerates differences resulting from radically changing the position of a single language in the tree,[2] it nevertheless provides a firm enough evaluation to squarely contradict the claim of Dunn et al. (2005:2074) that there is a "close match" between their typologically based Western Oceanic tree and the tree based on the comparative method or the similar claim in Dunn et al. (2007).

Dunn et al. (2007) state that they "devoted particular attention to" issues of areality in their (2005) paper, and this is easily verified; they noted that "[t]he results show a remarkably geographically consistent pattern: [t]he major clades represent archipelagos, and within each archipelago the nearest neighbors tend to form sister clades", that "regional diffusion also may account for the phylogenetic signal observed", and that "[a] second possibility is the null hypothesis of no relatedness between the Papuan languages. In that case, we would expect the orderly and geographically consistent *phylogenetic* signal that does emerge from the data" (2005:2074; emphasis ours). We agree; Dunn et al. have succeeded in modeling the spatial distribution of languages, such that the closer two languages are, the more similar they appear. We have not seen any indication that there *is* a phylogenetic signal, in the comparative linguistic sense. All indications are that an approach that uses exclusively typological data is highly sensitive to effects of diffusion; indeed, Dunn et al. (2007:397–98) note that "the evidence given in Dunn et al. (2005) hinged on the seemingly greater than chance congruence of the phylogenetic tree [the tree produced using their typology-based methods—DW&A] to geographic distribution." This seems to be a clear admission that the trees pick up geographic signals, as would be

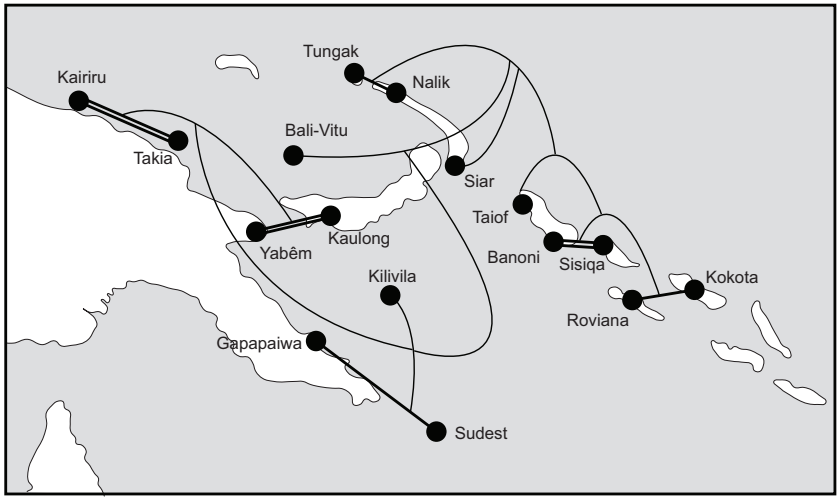
2. In particular, altering the position of a single node from one end of the tree to another will have drastic consequences for the RF measurement. This is not the case in the trees examined here, where there are no nodes that swap radically (compare the two representations in figure 2). Further, we avoided the problem of having too many trees in the generated set showing no match at all, probably because we did not use randomly generated trees; the 10,000 generated trees were constrained by a realistic branching model.

predicted by what we know about the diffusion of typological traits (e.g., Holman et al. 2007). The role of geography can be seen in figure 3.

In order to test the robustness of Dunn et al.'s technique it would need to be tested against a well-established phylogeny in which lower-level subgrouping does not correspond to geographic proximity. The case of Romanian comes to mind, as an example of a language belonging to the Romance subgroup of Indo-European that is geographically closer to the surrounding Slavic languages than to other members of its own subgroup. We have not coded up the features used in Dunn et al. (2005) for a sample of Romance and Slavic languages, but we do have available a classification based on phonological features. We coded 86 features representing the complete segmental phonologies of 33 languages, split between the Germanic, Romance, and Slavic sub-families of Indo-European. We used Sardinian as an outgroup (external taxon). The data was subjected to the neighbor-joining algorithm as implemented in SplitsTree (Huson and Bryant 2006), with the resulting tree shown as figure 4.[3] While (most of) Germanic forms a distinct branch of the tree, and Romance occupies the middle ground between Germanic and Slavic (while being grouped more closely with its western European neighbor, Germanic, suggesting areal effects), it is also clear that Romanian, on the basis of the structural features we encoded, is treated as a Slavic (or "eastern European") language (albeit an atypical one).

Other examples could be presented based on the *World atlas of language structures* (Haspelmath et al. 2005), for instance. We provide one in figure 5; here we selected sixteen of the better-documented languages in the database, used all the features attested for each language, and again produced the tree by neighbor-joining. While this tree does

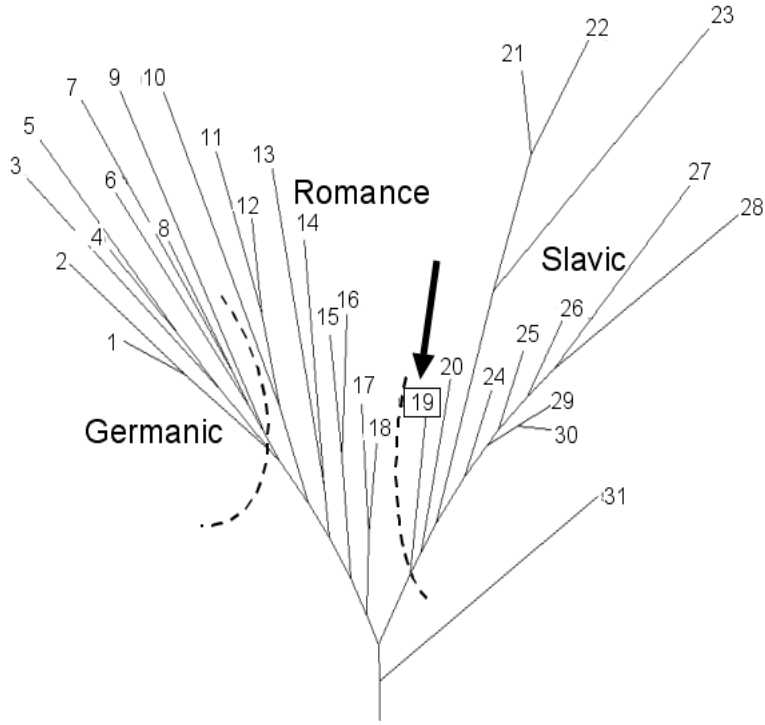
FIGURE 3. LINKS BETWEEN THE WESTERN OCEANIC LANGUAGES †



† Single lines show links from Dunn et al. (2005) that approximately match the analysis in Ross (1988). Links that are counter to Ross (1988) are shown as double lines (===), and in all cases join geographically close languages.

3. The features and coding are available at http://email.eva.mpg.de/~wichmann/DonohueWichmannAlbu2008-SupportingMaterials.pdf

FIGURE 4. INDO-EUROPEAN CLASSIFICATION USING PHONOLOGICAL FEATURES: ROMANIAN AS A SLAVIC LANGUAGE

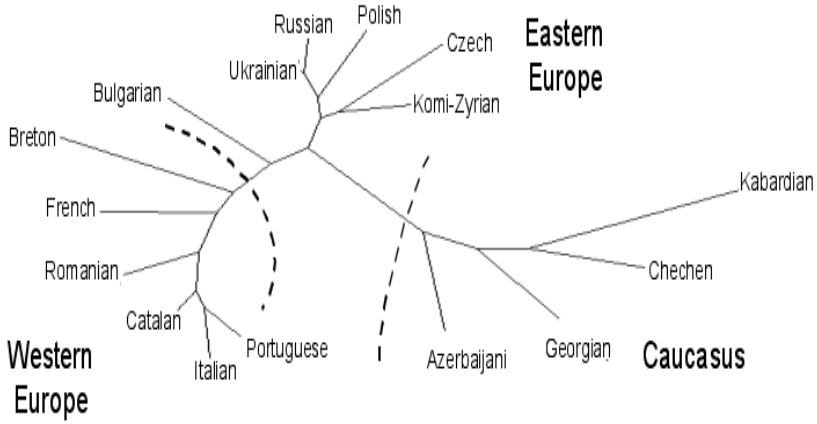


Key: 1 Letzebuergesch; 2 German; 3 Swiss German; 4 Dutch; Afrikaans; 6 Swedish; 7 Norwegian; 8 Frisian; 9 Danish; 10 Icelandic; 11 Galician; 12 Spanish; 13 Portuguese; 14 Catalan; 15 French; 16 Jerriais; 17 Italian; 18 East Lombard; 19 Romanian; 20 Polish; 21 Russian; 22 Bulgarian; 23 Sorbian; 24 Slovene; 25 Macedonian; 26 Czech, Slovak; 27 Ukrainian; 28 Belarussian; 29 Serbian, Croatian; 30 Bosnian; 31 Sardinian.

treat Romanian as a Romance language, the overall topology is determined by areal patterns. Thus, the Celtic language Breton, spoken in France, has French as its closest neighbor. The Uralic (Finnic) language Komi-Zyrian, spoken in northern Russia, is treated as a Slavic language, and at one end of this unrooted tree we see a clustering of four unrelated languages of the Caucasus region: Azerbaijani (Altaic, Turkic), Georgian (Kartvelian), Chechen (Nakh-Daghestanian), and Kabardian (North-West Caucasian).

The degree to which typological features may reveal genealogical rather than areal information may be expected to depend (to a certain extent) on the number of features employed (as well as their stabilities and the way in which they are encoded). The as-yet unpublished typological database Jazyki Mira, assembled under the auspices of the Russian Academy of Sciences and described in Polyakov and Solovyev (2006), is designed for the encoding of 3821 linguistic features and contains data for 315 Eurasian languages. Ongoing work in collaboration with Vladimir Polyakov and Valery Solovyev, to be reported more fully in future publications, has shown that these data, not surprisingly,

FIGURE 5. SOME LANGUAGES OF EURASIA CLASSIFIED USING TYPOLOGICAL FEATURES FROM HASPELMATH ET AL. (2005)



allow for establishing much more precise phylogenies than do the data of Haspelmath et al. (2005). Nonetheless, as will also be reported elsewhere, a lexicostatistical analysis based on a 40-item subset of the Swadesh list (selected for stability, as described in Holman et al. forthcoming) performs still better, to judge from comparisons of performance on a balanced sample of 39 languages. Thus, there is not currently any evidence to sustain the hope that even large amounts of typological features may somehow reveal phylogenetic signals that cannot be better revealed by more well-established methods in historical linguistics. However, because such features are prone to revealing areal patterns, including areal patterns that may sometimes have a considerable antiquity, they are more useful than, for instance, basic vocabulary if the aim of a given investigation is the study of language contact.

5. CONCLUSION. We have shown that clustering languages on the basis of typological features produces clusters that correlate most closely to geography, even when the relevant subgroups (as determined by the comparative method) do not follow geography. Thus, we agree with the consensus position of historical linguists, that typological data cannot better reveal genealogical relations among languages than more traditional methods in historical linguistics—a position which can be supported by an argument from probability (see Nichols 1996).

Better results may be achieved by enlarging databases, paying more attention to optimal ways of encoding features, selecting features based on their relative stabilities, and using optimal phylogenetic algorithms (Wichmann and Saunders 2007). Nevertheless, there is presently no evidence to sustain the hope expressed by Dunn et al. (2007:388) that their typologically based techniques "might apply where the comparative method cannot." This does not mean that typological features are uninteresting for historical lin-

guistics, however. They reveal some information on both genealogical and areal relations among languages, and the use of datasets that are amenable to statistical analyses makes it possible to evaluate results better than for non-quantitative data. Thus, we welcome the attempt of Dunn et al. (2005) to establish genealogies based on typological data as an interesting experiment, even if we differ from the authors in our assessment of the results of this experiment.

REFERENCES

Chu, Johan, and Christoph Adami. 1999. A simple explanation for taxon abundance patterns. *Proceedings of the National Academy of Sciences of the United States of America* 96:15017–19.

Donohue, Mark, and Simon Musgrave. 2007. Typology and the linguistic macro-history of island Melanesia. *Oceanic Linguistics* 46:348–87.

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

Dunn, Michael, Robert Foley, Stephen Levinson, Ger Reesink, and Angela Terrill. 2007. Statistical reasoning in the evaluation of typological diversity in Island Melanesia. *Oceanic Linguistics* 46:388–403.

Haspelmath, Martin, Matthew Dryer, David Gil, and Bernard Comrie (eds.). 2005. *The world atlas of language structures*. Oxford: Oxford University Press.

Holman, Eric W. 2008. Approximately independent features of languages. *International Journal of Modern Physics* C 19 (2):215–20.

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.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Vilupillai, André Müller, Pamela Brown, and Dik Bakker. Forthcoming. Explorations in automated lexicostatistics. *Folia Linguistica*.

Huson, D. H., and D. 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 Press.

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

Polyakov, Vladimir N., and Valery D. Solovyev. 2006. *Kompjuternye modeli i metody v tipologii i komparativistike* (Computer models and methods in typology and comparative linguistics). Kazan: Kazanskiy Gosudarstvennyy Universitet.

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

———. 2001a. Is there an East Papuan phylum? Evidence from pronouns. In *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.

———. 2001b. Contact-induced change in Oceanic languages in north-west Melanesia. In *Areal diffusion and genetic inheritance: problems in comparative linguistics*, ed. by Alexandra Aikhenvald and R. M. W. Dixon, 134–66. Oxford: Oxford University Press.

Wichmann, Søren. 2005. On the power-law distribution of language family sizes. *Journal of Linguistics* 41:117–31.

Wichmann, Søren, and Arpiar Saunders. 2007. How to use typological databases in historical linguistic research. *Diachronica* 24 (2):373–404.

mark@donohue.cc wichmann@eva.mpg.de mihai@evol.biology.mcmaster.ca