

Re: New Methodology on Analysis of Language Change

Source: <http://sci.tech-archive.net/Archive/sci.lang/2006-01/msg01204.html>

- *From:* "Joseph W. Murphy" <jwmurphy700@xxxxxxxxxxxxxxxx>
 - *Date:* Thu, 12 Jan 2006 13:02:07 GMT
-

Joe Murphy wrote:

>> I tried to figure out exactly what they were doing but failed. The quest
>> took me to the "Science" website and I have to pay \$10.00 for a day pass to
>> read the bigger article.

And On Wed, 11 Jan 2006 23:49:57 GMT, Peter T. Daniels wrote:

> I hope you didn't do so ... Your central public library, and even the
> bigger branches, should have *Science*; if it's current, you can buy it
> at Borders (or B&N).

And Joe Murphy now writes:

Eureka! Now "Science" is showing what looks like the entire article.
That wasn't the case a few days ago when I first looked at it.

It's here:

Structural Phylogenetics and the Reconstruction of Ancient Language History

Michael Dunn,^{1*} Angela Terrill,^{1,2} Ger Reesink,^{1,2} Robert A. Foley,³
Stephen C. Levinson^{1,2}

The contribution of language history to the study of the early dispersals of modern humans throughout the Old World has been limited by the shallow time depth (about 8000 ± 2000 years) of current linguistic methods. Here it is shown that the application of biological cladistic methods, not to vocabulary (as has been previously tried) but to language structure (sound systems and grammar), may extend the time depths at which language data can be used. The method was tested against well-understood families of Oceanic Austronesian languages, then applied to the Papuan languages of Island Melanesia, a group of hitherto unrelatable isolates. Papuan languages show an archipelago-based phylogenetic signal that is consistent with the current geographical distribution of languages. The most plausible hypothesis to explain this result is the divergence of the Papuan languages from a common ancestral stock, as part of late Pleistocene dispersals.

1 Max Planck Institute for Psycholinguistics, Post Office Box 310, 6500 AH Nijmegen, Netherlands.

2 Center for Language Studies, Radboud University, Post Office Box 9102, 6500 HC Nijmegen, Netherlands.

3 Leverhulme Centre for Human Evolutionary Studies, University of Cambridge, Downing Street, Cambridge CB2 3DZ, UK.

* To whom correspondence should be addressed. E-mail: michael.dunn@xxxxxx

The linguistic comparative method used to construct language family trees relies on recognizing "cognate sets": words in different languages that are related in meaning and form because they can be shown to have the same ultimate source in an ancestor language. The comparative method has helped define the major linguistic family groups that are recognized today. Unfortunately, because of the continual process of linguistic change, the method is limited to a time depth of approximately 8000 ± 2000 years (1). However, it is probable that a considerable portion of linguistic diversification occurred at earlier dates, associated with later Pleistocene human dispersals. Alternative attempts to reach further back and link the world's 300 language families (2) into larger taxonomic units are controversial (3).

One example of this older diversification may be found in Island Melanesia. Radiocarbon dating for Island Melanesia has demonstrated Pleistocene occupation more than 35,000 years ago (6, 7) (Fig. 1). Evidence suggests high levels of inter- and intrapopulational genetic variation (8, 9), with no simple relationship with linguistic patterns. The languages spoken in the area are of two groups: (i) over 100 languages belonging to four groups of the well-established Austronesian family, which probably originated in the area close to Taiwan and spread to this region about 4000 years ago (10); and (ii) 23 "Papuan" languages, which are not known to have any phylogenetic relation to one another and are of much greater antiquity in the region

The lexical evidence for relationships between Papuan languages is minimal. Apart from shared Austronesian loans, there are few plausible cognate candidates found in comparisons of pairs of words from Papuan vocabularies (Fig. 2) [see, however, (11)]. Assuming that the rate of vocabulary loss in the Papuan languages is similar to rates observed elsewhere, these languages are either unrelated or have been separated at least since the early Holocene or late Pleistocene. These languages do, however, show a high degree of structural similarity, distinguishing them as a group from their Austronesian neighbors, which has led scholars to propose genealogical (or near-genealogical) groupings (12, 13). In the absence of identifiable lexical cognates, we have used computational cladistic analysis of these features of linguistic structure to test whether a phylogenetic signal can be identified beyond the resolution of lexical form-based methods [for other cladistic methods using lexicons, see (14)]. The structural features of a language, like the lexicon, are subject to processes of decay over time and can also be borrowed or exchanged across languages. However, such exchange usually only occurs

under special conditions of prolonged and intensive contact, and it is at least plausible that where the lexical signal has been lost, a faint structural signal might still be discernible. Linguistic structure that is, grammar rather than vocabulary has previously been used in historical linguistics to show statistical evidence for ancient links between languages from different parts of the world (1, 2, 22, 23) but not directly to reconstruct phylogenetic relationships.

A questionnaire-based database was constructed, in which linguistic structural features were coded for their presence or absence in each of the target languages. These characters were abstract (coded without respect to their formal expression) and were selected to provide broad typological coverage, reflecting the known linguistic variation of the region (24), as well as to be features that would typically be described in a published sketch grammar. Traits invariant in the region (either entirely absent, such as polysynthesis or proximate/obviative case distinctions; or present in all the languages, such as the existence of a word class of verbs) were not coded. Characters that show strong implicational correlations were excluded, although characters with weaker tendencies to covariance were not excluded where the current state of linguistic typological knowledge does not allow us to systematically distinguish functionally motivated covariance from phylogenetic or areal patterns. The completed data matrix contained 125 binary features coded for 15 Papuan and 16 Austronesian languages spoken in an overlapping region. The Papuan database was mostly compiled by linguists with field experience in the language and was supplemented from published and unpublished sources where available. The Austronesian database was constructed from published sources (25). All sets of data were checked by a second coder to ensure consistency.

The binary-coded linguistic features allowed us to treat these as character traits distributed among taxonomic units (languages) and thus to apply cladistic algorithms (maximum parsimony or NeighborNet) to determine potential phylogenetic relationships among them (26).

The hypothesis that grammatical structure retained a phylogenetic signature was first tested among 16 languages belonging to the Meso-Melanesian, Papuan Tip, and North New Guinea linkages, three sister clades within the Western Oceanic subgroup of Austronesian, the relationship of which has been established by the comparative method (10, 27) {although not completely unambiguously, because there is lexical evidence in particular that the Papuan Tip and the North New Guinea linkages had a period of shared history after their separation from Meso-Melanesian [(10), p. 101]}. We carried out a parsimony analysis on the structural data from these languages, from which we obtained a consensus tree [tree length, 224 steps; consistency index (CI) = 0.42; rescaled consistency index (RC) = 0.19; retention index (RI) = 0.46]. When this tree (Fig. 3, right) is compared with the classification based on the comparative method (Fig. 3, left), there is a close match. In the consensus tree, the Meso-Melanesian group forms a major branch. Papuan Tip and North New Guinea together form a clade, with the North New Guinea linkage nested as a subclade within it. This is consistent with uncertainties in the linguistic reconstruction. The

internal structure of the Meso–Melanesian group is quite flat, but all except one of the clades posited by the comparative method are congruently represented in the consensus tree. These results show that cladistically analyzed grammatical structure can preserve a signal that is consistent with a known phylogeny derived by traditional lexical techniques.

On the basis of this result, we applied the same method to a set of languages in which lexical similarities are not present. Taking 15 Papuan languages for which we have full structural data and applying the same methods, we obtained a consensus tree of the most parsimonious cladograms for the bootstrapped data set (Fig. 4). This tree has a tree length of 349 steps, CI = 0.35, RC = 0.14, and RI = 0.39. The results show a remarkably geographically consistent pattern: The major clades represent archipelagos, and within each archipelago nearest neighbors tend to form sister clades, despite a nearly complete absence of lexical relatedness.

Interpretation is problematic, because there are no generally accepted independent linguistic criteria for assessing the Papuan trees. One possibility is that these trees reflect contact with local Austronesian neighbors, providing an areal rather than phylogenetic signal. In experiments, combined Austronesian–Papuan consensus trees were in some cases intermeshed, but the result was statistically weak (28). Because Papuan and Austronesian are very unlikely to be genuine sister clades, a high degree of homoplasy can be the result of either contact or chance convergence, and combined trees of very remotely related families are likely to be less robust than those where there are good grounds for assuming monophyly. A second possibility is the null hypothesis of no relatedness between the Papuan languages. In that case, we would not expect the orderly and geographically consistent phylogenetic signal that does emerge from the data. This signal is consistent with migration followed by divergence through local isolation. A further possibility is that the geographically consistent tree reflects recent areal contact among Papuan speakers, but most of these languages are not currently spoken in contiguous regions. Because these languages may have been contiguous in the past, regional diffusion also may account for the phylogenetic signal observed, a possibility that we cannot test without more detailed archaeological information.

We therefore suggest that this method reveals evidence of large–scale genealogical clustering of the Island Melanesian languages; the lack of putative lexical cognates dates these relationships considerably before the Austronesian arrival, in line with the radiocarbon dates from the later Pleistocene, when humans entered Island Melanesia from mainland Papua New Guinea.

There remain important issues to resolve. The first is methodological; bootstrap values, especially in the deeper branches, are low by comparison with biological systems, and further work is required to determine whether this reflects rates of convergence, trait covariation, or processes other than phylogenesis alone. Second, the branching sequence does not fit the generally expected dispersal path. A priori, Island Melanesian Papuan

Re: New Methodology on Analysis of Language Change

languages should show a general west-to-east pattern of diversification, with the center of diversity in the west. The results of our data are more complex. In particular, the position of the Solomons languages is anomalous, located in the tree between the Bismarcks clade and the Bougainville clade, in violation of geographic expectation [because Bougainville is the natural way-station on the route from mainland New Guinea to the Solomons (Fig. 1)]. During the late Pleistocene, Bougainville and the Solomons were united into a single island, from which the Bismarcks were always separate. A plausible interpretation of the Papuan language tree is thus that the two language groups now located on the Solomons and Bougainville separated from a common ancestor. This could have happened while they could still freely migrate on a common landmass, a time depth (10,000 years) in accord with that required to erode traces of common vocabulary. This population history hypothesis will require further testing with both linguistic and genetic data.

If grammatical structures can retain a phylogenetic signal beyond the current temporal ceiling on the reconstruction of language history, then the possibility is opened up of finding relationships between others of the world's 300 or so existing language families and isolates.

References and Notes

1. J. Nichols, *Linguistic Diversity in Space and Time* (Univ. of Chicago Press, Chicago, 1992).
2. J. Nichols, in *The Handbook of Historical Linguistics*, B. D. Joseph, R. D. Janda, Eds. (Blackwell, Oxford, 2003), pp. 283-10.
3. J. H. Greenberg, *Language in the Americas* (Stanford Univ. Press, Stanford, CA, 1987).
4. L. L. Cavalli-Sforza, P. Menozzi, A. Piazza, *The History and Geography of Human Genes* (Princeton Univ. Press, Princeton, NJ, 1994).
5. D. Bolnick, B. Shook, L. Campbell, I. Goddard, *Am. J. Hum. Genet.* 75, 519 (2004). [CrossRef] [ISI] [Medline]
6. P. Kirch, *The Lapita Peoples* (Blackwell, London, 1997).
7. M. Spriggs, *The Island Melanesians* (Blackwell, London, 1997).
8. D. A. Merriwether et al., in *Genomic Diversity: Applications in Human Population Genetics*, S. S. Papiha, R. Deka, R. Chakraborty, Eds. (Kluwer Academic/Plenum, New York, 1999), p. 153.
9. M. Kayser et al., *Am. J. Hum. Genet.* 72, 281 (2003). [CrossRef] [ISI] [Medline]
10. J. Lynch, M. Ross, T. Crowley, *The Oceanic Languages* (Curzon Press, Richmond, UK, 2002).
11. M. Ross, in *The Boy from Bundaberg: Studies in Melanesian Linguistics in Honour of Tom Dutton*, A. Pawley, M. Ross, D. Tryon, Eds. (Pacific Linguistics, Canberra, Australia, 2001), pp. 301-21.
12. S. A. Wurm, *Papuan Languages of Oceania* (Gunter Narr Verlag, Tübingen, Germany, 1982).
13. J. H. Greenberg, in *Current Trends in Linguistics*, vol. 8, *Linguistics in Oceania*, T. A. Sebeok, Ed. (Mouton and Co., the Hague, 1971), pp.

807 871.

14. D. Ringe, T. Warnow, A. Taylor, *Trans. Philol. Soc.* 100, 59 (2002).
[CrossRef] [ISI]
15. P. Forster, A. Toth, *Proc. Natl. Acad. Sci. U.S.A.* 100, 9079 (2003). [Abstract/Free Full Text]
16. R. D. Gray, Q. D. Atkinson, *Nature* 426, 435 (2003). [CrossRef] [ISI] [Medline]
17. R. D. Gray, F. M. Jordan, *Nature* 405, 1052 (2000). [CrossRef] [ISI] [Medline]
18. C. J. Holden, R. Mace, *Proc. R. Soc. London Ser. B* 270, 2425 (2003). [CrossRef] [ISI]
19. C. J. Holden, *Proc. R. Soc. London Ser. B* 269, 793 (2001). [ISI]
20. K. Rexová, D. Frynta, D. J. Zrzav, *Cladistics* 19, 120 (2003). [CrossRef] [ISI]
21. A. McMahon, R. McMahon, *Trans. Philol. Soc.* 101, 7 (2003). [CrossRef] [ISI]
22. J. Nichols, in *Historical Linguistics 1993. Papers from the Eleventh International Conference on Historical Linguistics*, H. Andersen, Ed. (John Benjamins, Amsterdam, 1995), pp. 337-56.
23. J. Nichols, in *The Origin and Diversification of Language*, N. Jablonski, L. C. Aiello, Eds. (California Academy of Sciences, San Francisco, 1998), pp. 127-70.
24. M. Dunn, G. Reesink, A. Terrill, *Ocean. Linguist.* 41, 28 (2002). [ISI]
25. See sources of language data in the supporting online material.
26. Materials and methods and a description of linguistic characters are available as supporting material on Science Online.
27. M. Ross, *Proto Oceanic and the Austronesian Languages of Western Melanesia* (Pacific Linguistics C-98, Canberra, Australia, 1988).
28. This intermeshing of trees does reflect long-term contact in New Britain (26).
29. D. Tryon, B. Hackman, *Solomon Island Languages: An Internal Classification* (Pacific Linguistics C-72, Canberra, Australia, 1983).
30. This work, as part of the European Science Foundation EUROCORES Programme OMLL, 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. Additional fieldwork data used in this study were provided by E. Lindström, S. Robinson, T. Stebbins, W. Thurston, and C. Wegener; assistance with coding from published sources was provided by S. Nordhoff and V. Rodrigues. We thank M. Mirazon-Lahr, E. Lindström, and G. Senft for discussion.

There are some illustrations (graphs and tree charts) that are in the article that I didn't try to reproduce. Also there is a "content" page in PDF format that can be read online here:

<http://www.sciencemag.org/cgi/content/full/309/5743/2072/DC1>

Joe Murphy
Boy Linguist

.

- **Follow-Ups:**
 - ◆ **Re: New Methodology on Analysis of Language Change**
 - ◇ From: Peter T. Daniels

- **References:**
 - ◆ **New Methodology on Analysis of Language Change**
 - ◇ From: Joseph W. Murphy
 - ◆ **Re: New Methodology on Analysis of Language Change**
 - ◇ From: Peter T. Daniels
 - ◆ **Re: New Methodology on Analysis of Language Change**
 - ◇ From: Joseph W. Murphy
 - ◆ **Re: New Methodology on Analysis of Language Change**
 - ◇ From: Peter T. Daniels

- Prev by Date: **Re: New Methodology on Analysis of Language Change**
- Next by Date: **Re: New Methodology on Analysis of Language Change**
- Previous by thread: **Re: New Methodology on Analysis of Language Change**
- Next by thread: **Re: New Methodology on Analysis of Language Change**
- Index(es):
 - ◆ **Date**
 - ◆ **Thread**