Consensus and the lexicon in historical linguistics

Rejoinder to "Basic vocabulary and Bayesian phylolinguistics"

Mark Donohue, Tim Denham and Stephen Oppenheimer
The Australian National University / La Trobe University / University of Oxford

1. Introduction

We are grateful to Greenhill & Gray for their comments and the opportunity to engage in constructive debate on computational phylogenetic methods. We first reiterate our main positions:

1. We are excited by the prospects for understanding linguistic and social histories made available by the ability to deal with large bodies of data with the impartiality that many different computational algorithms offer. We note, e.g., the general success of the ASJP program in replicating Indo-European subgroups on wordlists of only 40 items (e.g. Wichmann et al. 2010).
2. We do not agree with the assumption that lexical data is best examined in a model that assumes relatedness is only due to descent. When linguistic subgroupings based on the comparative method do not correspond to those derived from Gray et al.'s methods (or other lexicon-based methods), the computational model reflects social geography (contact, or areality), rather than family relationships.
3. New methodologies must be calibrated against languages and subgroups with well-established histories, so that researchers can assess how well they perform.
4. The comparative method is the best means to infer ancestry for any particular language and any linguistic phylogeny.

Diachronica 29:4 (2012), **538–546**. DOI 10.1075/dia.29.4.06don
ISSN 0176–4225 / E-ISSN 1569–9714

We advance the following responses to the seven concerns that Greenhill & Gray raise in their reply:

1. Innovations are not properly utilised or distinguished; we address this point below.
2. Sound change is not utilised, as explained in the article.
3. Lexical items are frequently borrowed, as discussed in the sources cited in our original article and confirmed by Greenhill & Gray's statements in their supplementary materials.[1]
4. There are discrepancies in the placement of individual languages, as Greenhill & Gray note. We did not raise this point, but agree.
5. Greenhill & Gray consistently fail to group Tongan and Niuean together, the two languages that form a first-order subgroup opposed to the rest of Polynesian, making their Polynesian tree incompatible, at the highest level, with *all* informed linguists' opinions.
6. Regardless of the complexity of trees, evaluating replication of subgroups (the aim of the exercise) is simple. The quantifiable dominance of social geography when not congruent with (linguists') trees suggests that 'quartets distance' is a weak measure of the successful replication of linguistic hierarchies. We illustrate with the case of Malayo-Chamic below.
7. Greenhill & Gray's trees show the effects of contact at numerous points where geography and phylogeny do not coincide, as they acknowledge with such phrasing as "these differences could be explained by unidentified borrowings" (supplement).

For further discussion of concerns 3, 4 and 7, see the supplementary materials to this rejoinder. Point 5 is discussed at length in our article, and points 1, 2 and 6 are discussed here.

1. Greenhill & Gray ask whether we claim 'that rates of borrowing in basic vocabulary are at least as high as rates of inheritance'. We did not make this claim, and the short answer is that we do not know, not having conducted the study. We point out that (a) 'basic' vocabulary is culture-dependent, and not unproblematic (see work reported in Wichmann et al. 2010), and (b) the wordlist used by Gray et al. is a mixture of kinds of vocabulary, including items (such as "intestines", "sew", "spider", among others) that are not generally included as basic, and which are (in our experience) difficult to elicit consistently. Since we *do* have the attested example of Jarai, with 48% identified loans, we are confident to state that 'rates of lexical borrowing *can be* as high as rates of inheritance'. The problem of identifying loans within the same cognate class, if there are no telling differences in sound change, and in the absence of epigraphic data, has not been addressed, yet should not be understated.

Replication

The term 'replication' has caused some confusion. 'Replication' is a term adopted from information theory and composed of the twin notions of *precision* and *recall* (van Rijsbergen 1979). *Precision* measures the exclusion of undesired elements and *recall* measures the inclusion of desired elements. Greenhill & Gray reply with phrases like '(not) invalidate' and 'support', apparently assuming that these are equivalent to replication, which they are not.

We illustrate the differences in usage with reference to Malayo-Chamic, which Greenhill & Gray believe us to have misinterpreted. Greenhill & Gray state that "our results do group the Malayic and Chamic languages into a higher-order subgroup with moderate posterior probability (0.60). Inside this group there are other languages that do not belong to Malayo-Chamic (e.g. Javanese), but this does not invalidate that group" (online supplement). It appears that 'does not invalidate' applies when a targeted subgroup can be recalled somewhere in their tree, regardless of the precision of the clade required to recall it. This is clear when we compare Adelaar's (2005) comparative method-derived topology of the Malayo-Sumbawan group (Figure 1) with the clade shown in Gray et al. (2009, Figure 2). In both figures the Malayic and Chamic languages, together with the Bali-Sasak-Sumbawa group, with which they form a subgroup, are outlined. The number of languages in the tree from Gray et al. (2009) is shown under the labels in Figure 2, including the seven "other languages that do not belong", from four different, neighbouring subgroups (in bold in Figure 2). The position of Balinese, Madurese, Sasak and Sundanese in Figure 2 is at variance to Adelaar's tree in Figure 1; the only node that contains all the Malayo-Chamic languages is, in fact, the Malayo-Sumbawan node. The North Sumatran and Lampungic groups, as well as Javanese and Rejang, form subgroups separate from Malayo-Sumbawan in both Adelaar's and Blust's classification.

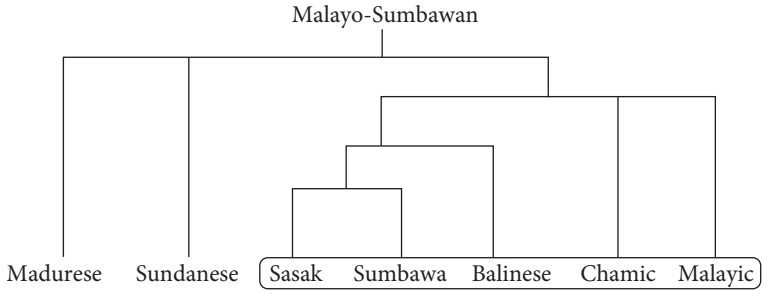

Figure 1. Malayo-Sumbawan relations (Adelaar 2005).

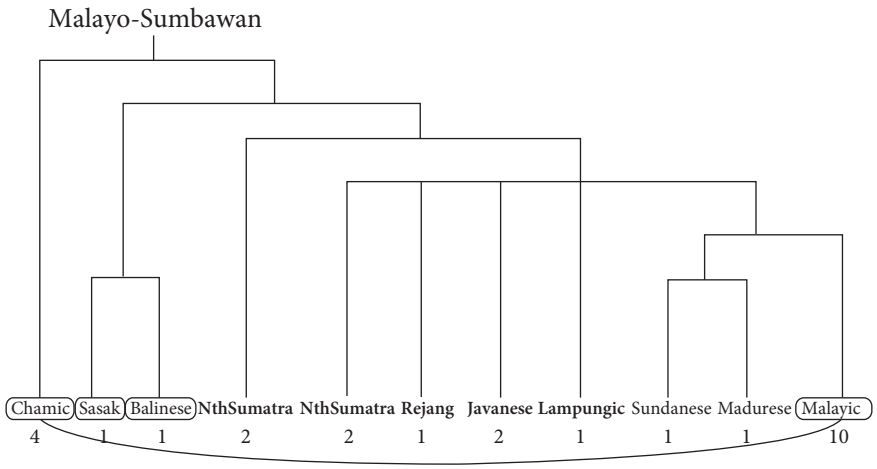

Figure 2. Malayo-Sumbawan relations (Gray et al. 2009).[2]

Greenhill & Gray state that their clustering 'does not invalidate' Malayo-Sumbawan. We can understand the basis of their claim, but only with the understanding that *recall* is what they are seeking to achieve without *precision*.[3] If precision is not an important factor in evaluating language subgroups, any subgroup can be recalled by simply reproducing the entire list of languages in the survey. For this reason it is essential to use the information structure measures of replication, incorporating recall and precision, rather than the methods that Greenhill & Gray mention.

Innovations versus retentions

Greenhill & Gray explain how their method differentiates innovations and retentions in the choice of lexical items associated with particular semantic concepts. It is unfortunate that the example they offer for the lexeme "sun" relies on positing a new subgroup, 'the Polynesian languages plus Rotuman', that is not recognised by any linguists.

2. Note Gray et al.'s method places some languages of the North Sumatra group in two different positions inside Malayo-Sumbawan. Members of this subgroup also appear in other positions outside the Malayo-Sumbawan group shown here.

3. A quantified assessment of the degree of fit of Gray et al.'s (2009) tree to Adelaar's (2005) tree is shown in the supplementary materials to this rejoinder. The whole Malayo-Sumbawan group is replicated at a 0.84 level; exactly the rate displayed in our original Figure 3, which represents a poor clustering of European languages, one that Greenhill & Gray describe as one that has "some notable misplacements". Gray et al.'s (2009) replication of Malayo-Chamic is considerably worse (0.72); see the supplementary materials.

We refer, however, to the problem of sound change and innovation, which, given what we know of naturalness and sound change, no amount of automated decision-making (as invoked by Greenhill & Gray) will resolve (the problem is only magnified if we consider the effects of mergers). Greenhill & Gray ask "Is there any sense in which we neglect sound changes?", and answer that "we decided not to model sound change *within* cognates". The loss of information this entails was illustrated with the discussion of Polynesian data in Table 1 of our original article. We make the point more explicitly with the data in Table 1, showing reflexes of the Austronesian reconstruction *qijuŋ "nose" in a representative selection of Central Pacific languages (all different forms for Nuclear Polynesian are listed, and though only one language per form is indicated all are attested more than once). It would be hard not to reconstruct *isu for the morpheme at a Proto-Polynesian level. A stochastic method would, with high probability, assume *ihu, since the *s is preserved as such only in Nuclear Polynesian, whereas *h is found at all levels of the Polynesian tree. For reasons such as these we believe that their method fails to adequately take into account innovations and retentions.

Table 1. Innovations and retentions in a cognate morpheme: Central Pacific "nose".

Subgroup			**"nose"**
Fijian			iðu
Rotuman			iʔu
Polynesian	Tongic	Tongan	ihu
		Niuean	ihu
	Nuclear Polynesian	Anuta etc.	iu
		Emae etc.	isu
		Hawaiian etc.	ihu
		Mangareva etc.	iʔu

3. General issues

Greenhill & Gray state that:

> None of the issues raised by Donohue et al. cast doubt on our central findings about the origin, age, expansion sequence and manner of the spread of Austronesian languages. … quantification of the overall similarity between our trees and those of the comparative method reveals a high level of congruence.

The application of the comparative method unambiguously places the origin of the Austronesian languages in Taiwan (see Blust 2009 and Ross 2009 for recent

synopses). The age of the Austronesian family has been inferred on the basis of archaeological work (see Donohue & Denham 2010, with references). The 'expansion of the Austronesian language family' (= internal (hierarchical) subgrouping) was determined by the research of numerous historical linguists applying the comparative method (for overviews see Blust 2009, Ross 1995, 2009, and others). The notion of a series of pulses and pauses in the spread of Austronesian has a long history (cf. Blust's 2009 overview). Rather than claim that these historical inferences are 'findings' derived from the work of Gray et al. (2009), we might better state that by taking data coded following comparative method principles, and using Bayesian phylogenetic trees and various algorithms for fitting dates to nodes, Gray et al. (2009) derive a classification similar to the comparative method (but differing from it when geography is not congruent).

Concerning congruence, Greenhill & Gray state that 400 languages can be grouped in 5.8×10^{984} ways in a tree constrained to have only binary branches. Nonetheless, the tree that Gray et al. (2009) present, with a number of 100% probability nodes, fails (in numerous and dramatic ways) to match the tree against which they test their model (derived from Blust 2009). It may be that many of the 'traditional subgroupings' are 'not invalidated' by misplacements of languages; if, however, Armenian was placed within Indo-Iranian (to use their example), the result would still unsuccessfully reproduce Indo-European subgrouping.

There are methodological problems with Greenhill & Gray's claims for their method. The method relies on a number of non-automated judgments or tweaks. For instance, Gray et al. take advantage of the comparative method to identify cognate classes and then evaluate their results against comparative method derived trees:

1. Cognate decisions are made by historical linguists, using, as Greenhill & Gray point out, the evidence of regular sound correspondences to evaluate cognacy;
2. Loanwords are marked manually, as identified by the relevant linguistic experts; and
3. Archaeological dates are critical to their dating algorithm, yet the method through which archaeological findings and associated dates are matched to linguistic nodes in the tree is problematic given the lack of correspondence between linguistic and material cultural traits in parts of Island Southeast Asia, and the lack of any epigraphic findings of the relevant antiquity. A dating based only on *linguistic* data that matched the archaeological dates would be a genuine triangulation, and this has not (to our knowledge) been presented in any of their publications.

Given the incorporation of expert linguistic judgments and work — primarily derived from the comparative method — that went into the coding of data before it could be used, Gray et al.'s outcomes should be more covariate with the phylogeny

derived from the comparative method. Greenhill & Gray explain the method as necessarily involving expert judgments of cognacy — assessed after the historical phonology has been analysed and loans have been identified. However, their method discards the information about sound change, thereby losing some of the most valuable data that should be employed. This is not the case in other computational work, such as Nakhleh, Ringe & Warnow (2005).

Based on their classification of well-studied families such as Indo-European and Austronesian, it seems that Gray et al.'s (2009) method shows results similar to those of the ASJP program (which has been applied to classifications of many language families in Wichmann et al. 2010; software and data available at http://email.eva.mpg.de/~wichmann/software.htm) (Pompei et al. 2011). Unlike Gray et al.'s method, the ASJP method is entirely automated, serving as an aid to traditional historical methodologies. It is harder to justify a non-automated computational approach that uses the results of the comparative method to weakly emulate comparative method results.

Consequently, the 'striking congruence' Gray et al. (2009) and Greenhill & Gray claim is not based solely on unstructured computational modelling of lexical data. Given the nature of their data, Gray et al.'s method shows a strikingly *low* congruence with those developed by means of the comparative method.[4] Again, the number of possible trees is immaterial to the question, since Greenhill & Gray do not achieve a high degree of replication, due to low levels of precision when reproducing established subgroups.

The issue of loans

Greenhill & Gray write that they "used methods that are robust to the effects of borrowing", "spent considerable time identifying loanwords", and "did not include languages that were known to have high levels of borrowing". Citing figures from the World Loanword Database,[5] they claim that the lexemes used are on the low

4. Appealing to broad success with other language families begs the question of a comparative method derived tree for those families. We know of no such tree for Aslian, Bantu or Japonic, and note the controversy that surrounds Semitic. Consequently, claiming that a new method successfully replicates an existing subgrouping hypothesis is unfounded in the absence of a quantified calibration study (such as Wichmann et al. 2010). Similarly, Greenhill & Gray comparing their tree with a tree from Ethnologue, which contains many subgroups not supported by any linguistic work, does not demonstrate congruency "with the results of the comparative method".

5. Greenhill & Gray state that "The 41 languages in this sample are biased towards those known to have many loanwords". This is unusual, given that Mandarin Chinese is listed in *WOLD* as

end of the borrowability continuum, and that "we showed that the estimates of tree topology and time-depth were very robust despite quite high levels of borrowing between languages, around 20% of basic vocabulary every 1000 years". We are told in Greenhill & Gray's supplement that they have "uncovered 17 previously unrecognised loan words in the Sama-Bajaw language Inabaknon." If these additional 17 items (8%) "are the likely explanation for the minor mismatch between our results and the traditional linguistic subgroupings in this region" — where we would rephrase 'minor mismatch' as 'placement in the wrong (sub)group' — we fail to see how the two statements can be reconciled. If 8% undetected borrowing in a language previously identified with 3% loans can result in misplacement, it is not clear how 20% undetected borrowing every thousand years is not a problem.[6]

Conclusion

We are excited at the prospect of ever more sophisticated linguistic histories and chronologies emerging from the fusion of computational tools and linguistic expertise. At the same time, language histories are not simply records of inheritance, but also accounts of contact and convergence, just as societal histories involve more than just biology (e.g., Donohue & Denham 2011).

The two components of linguistic history, phylogeny (or inheritance derived from an original proto-language) and reticulation (or language flow from subsequent social interaction) are always entwined. The main point of our original paper is that Gray et al. (2009) and Greenhill & Gray consistently overestimate the phylogenetic contribution and consistently underestimate the effects of contact in their reconstruction of linguistic history, a problem magnified by their reliance on purely lexical data. Based on Greenhill & Gray's explication, Gray et al.'s (2009) method has a greater (indeed, great) contribution to make to our understanding of language change through social interaction than it does to understanding earlier dispersals from proto-languages.

having 1.2% loans. Our understanding is that the *WOLD* languages were selected on the basis of geographic distribution and the availability of experts.

6. Greenhill & Gray question how we can cite a figure of 48% loans in Jarai. We added the numbers in Grant (2005) and did the maths.

References not in main paper

Donohue, Mark & Tim Denham. 2010. "Farming and Language in Island Southeast Asia: Reframing Austronesian history". *Current Anthropology* 51.223–256.

Donohue, Mark & Tim Denham. 2011. "Language and Genes Attest Different Histories in Island Southeast Asia". *Oceanic Linguistics* 50.536–542.

Nakhleh, Luay, Don Ringe & Tandy Warnow. 2005. "Perfect Phylogenetic Networks: A new methodology for reconstructing the evolutionary history of natural languages". *Language* 81.382–420.

van Rijsbergen, C. J. Keith. 1979. *Information Retrieval*. London: Butterworth.

Ross, Malcolm D. 1995. "Some Current Issues in Austronesian Linguistics". *Comparative Austronesian Dictionary: An introduction to Austronesian studies* ed. by Darrell T. Tryon, Malcolm Ross, Adrian Clynes, Charles Grimes & Alexander Adelaar, Vol. 1, Pt. 1, 45–120 (= *Trends in Linguistics*, Documentation 10). Berlin: Mouton de Gruyter.

Ross, Malcolm. 2009. "Proto-Austronesian Verbal Morphology: A reappraisal". *Austronesian Historical Linguistics and Culture History: A festschrift for Robert Blust* ed. by Alexander Adelaar & Andrew Pawley, 295–326. Canberra: Pacific Linguistics.

Wichmann, Søren, Eric W. Holman, Dik Bakker & Cecil H. Brown. 2010. "Evaluating Linguistic Distance Measures". *Physica A* 289.3632–3639.

Authors' Addresses

Mark Donohue
Linguistics/School of Culture, History and Language
The Australian National University
Canberra, Australia ACT 0200

Tim Denham
Archaeology Program/School of Historical and European Studies
La Trobe University
Melbourne, Australia 3086

Stephen Oppenheimer
Institute of Cognitive and Evolutionary Anthropology/School of Anthropology
University of Oxford
64 Banbury Road
Oxford, United Kingdom OX2 6PN