Supplementary materials for "Consensus and the lexicon in historical linguistics": rejoinder to "Bayesian phylolinguistics"

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

We use these supplementary materials to systematically address the significant issues raised by Greenhill & Gray. Some of these have been addressed in the published rejoinder, and they are addressed here, more explicitly, or in more detail. Most concern misreadings of our article.

Precision, recall and the replication of subgroups

There appears to be some misunderstanding of our term 'replication',which Greenhill & Gray assume (from their response) to be equivalent to the phrase 'show support for'. Our usage of the term derives from information theory (Rijsbergen 1979), in which the term *replication* is composed of two measures, *precision* and *recall*. *Precision* is a measure of how accurately a desired cluster or grouping is achieved, excluding those members of the search space that are not required. *Recall* is the measure of how successfully the entire desired cluster is replicated. Put another way, *precision* measures how successfully undesired elements are excluded, while *recall* measures how successfully desired elements are included. Formally, both precision and recall are expressed on a scale from 0 to 1 and are calculated as shown in (1) and (2):

(1) Precision = (true positives) / (true positives + false positives)

(2) Recall = (true positives) / (true positives + false negatives)

These ideas can best be explained with reference to an example.

Suppose a new technique attempted to model the subgrouping of a number of Indo-European languages. The new clusters arising from the application of this hypothetical technique are identified by letters (A, B and C, or D and E) compared to the established subgroups, Germanic and Romance (Table S1).

Table S1. Comparing established subgroups with hypothetical clusters.

Language	Subgroup	New cluster	New cluster'
English	Germanic	A	D
Frisian	Germanic	A	D
Dutch	Germanic	A	D
German	Germanic	A	D
Swedish	Germanic	C	D
Norwegian	Germanic	B	D
Spanish	Romance	B	D
Catalan	Romance	B	D
French	Romance	B	D
Italian	Romance	A	E

The most generous interpretation of the new results is that cluster A approximates Germanic, and Cluster B replicates Romance. We can demonstrate this intuitive result by quantifying the extent to which each new cluster corresponds to an established subgroup. In Table S2 we show for each combination the precision and recall figures that result from the application of the formulas shown in (1) and (2). For instance, calculating the degree to which A replicates Germanic we note that of the five languages included in cluster B, four are in fact Germanic, yielding a precision value of 0.8 (4 true positives, divided by 4 true positives plus 1 false positive, Italian. The recall value is lower, 0.67, since the calculation here is the number of true positives (4) divided by the sum of the 4 true positives and the false negatives (2; Swedish and Norwegian).

Table S2. Calculating precision and recall in the data from Table S1.

		A	B	C
Germanic	Precision	= (4/(4+1)) = 0.8	= (1/(1+4)) = 0.25	= (1/(1+0)) = **1.0**
	Recall	= (4/(4+2)) = **0.67**	= (1/(1+5)) = 0.17	= (1/(1+5)) = 0.17
Romance	Precision	= (1/(1+4)) = 0.2	= (3/(3+1)) = **0.75**	= (0/(0+0)) = (0)
	Recall	= (1/(1+3)) = 0.25	= (3/(3+1)) = **0.75**	= (0/(0+4)) = 0.0

Note that we do not talk about 'invalidating' or 'showing support for' groups; we simply measure the fit to which the new clusters match the established subgroups. The reasons for this more nuanced pair of measures are simple: it is trivially simple to 'show support for' a group if all we consider is recall, or just precision. We have seen that the new cluster C has the highest precision score for the replication of Germanic, since the single-member cluster contains only one language, Swedish, and that one language is Germanic. The very low recall values, however, discount this as a valid replication. Similarly, examining the New cluster' column we can see that cluster D does a superlative job of supporting the Germanic subgroup: all Germanic languages in the test are found in this cluster, and so recall is 100%. Precision, however, falls to 0.67 (6/(6+3)), an unacceptably low level. The two metrics, precision and recall, can be combined in a harmonic mean, which we have done in our discussion of Polynesian in the main article. For comparision, the evenly-weighted harmonic means (of precision and recall values) for the replication of Germanic and Romance in the example presented above are: Germanic by A: 0.73; Romance by B: 0.75 (unsurprisingly); Germanic by C: 0.29. Cluster D has a harmonic mean of 0.80 for the replication of Germanic. Compare to Figure 3 in our original article, which replicates subgroups of Indo-European with a harmonic mean value of 0.84.

How is this relevant to the discussion of the claims debated here? Greenhill & Gray take issue with several of our objections, dismissing an initial three ('Bima, Malayo-Chamic, Paiwan') as 'mistakes from misreading the tree'. We take issue with the claim that we have misread their trees. In the published text of our rejoinder we have discussed Malayo-Chamic in detail, and here we will briefly discuss the other two groups following the quantification of the match mentioned in our published materials.

Quantifying the degree of fit of Gray et al.'s tree to Adelaar's tree is shown in Table S3. Here the true positives ("tp"), false positives ("fp") and false negatives ("fn") are shown for the

replication of a target, Malayo-Chamic or Malayo-Sumbawan, but different clusters in Gray et al.'s tree. Perfect precision is attained if only the Malayic or Chamic subgroup is considered; but in both cases recall suffers, resulting in a low combined value. Using Malayic, where the majority of the languages that Gray et al. consider, results in a higher replication rate for Malayo-Chamic – except that, of course, it is only replicating Malayic (and that perfectly), since without Chamic there is no Malayo-Chamic. If the whole Malayo-Sumbawan group is used ('our results do group the Malayic and Chamic languages into a higher-order' - supplement) recall is perfect, but precision drops to only just above 50%, due to the presence of seven languages from four 'intruder' subgroups. To give an appropriate example, the whole Malayo-Sumbawan group is replicated at an 84% level; exactly the rate displayed in our original Figure 3, which represents a clustering of European languages that Greenhill & Gray themselves describe as having 'some notable misplacements'.

Table S3. Replication rates for Malayo-Chamic and Malayo-Sumbawan.

Target subgroup:	Replicated by (cluster):	tp	fp	fn	Precision	Recall	Harmonic Mean
Malayo-Chamic	Malayic	10	0	4	1.0	0.71	0.83
Malayo-Chamic	Chamic	4	0	10	1.0	0.29	0.44
Malayo-Chamic	Malayo-Sumbawan	14	11	0	0.56	1.0	0.72
Malayo-Sumbawan	Malayo-Sumbawan	18	7	0	0.72	1.0	0.84

Greenhill & Gray claim that the splitting of Malayic and Chamic into two different locations across their Malayo-Sumbawan cluster, and adding in languages from four other subgroups 'does not invalidate that group'. This may be so, but if so we fail to understand what could 'invalidate' a group.

The second of our alleged 'misreadings' concerns the placement of Bima in Bima-Sumba. Gray et al. (2009, supplementary materials) state that there is 'a strongly supported clustering (1.00) of the controversial Bima-Sumba subgroup (0.82, Blust 2008)'. However, as support they cite the article in which Blust argues that the subgroup *does not exist*:

> The primary purpose of this paper is to show that there is strong evidence for a Sumba-Hawu group, and more restricted evidence for a larger subgroup that includes many or all languages of western and central Flores, but that Bimanese can be included in this group only if many of the languages of Esser's "Ambon-Timor" group are included in it as well. (Blust 2008: 48)

Greenhill & Gray acknowledge this, but subsequently claim that:

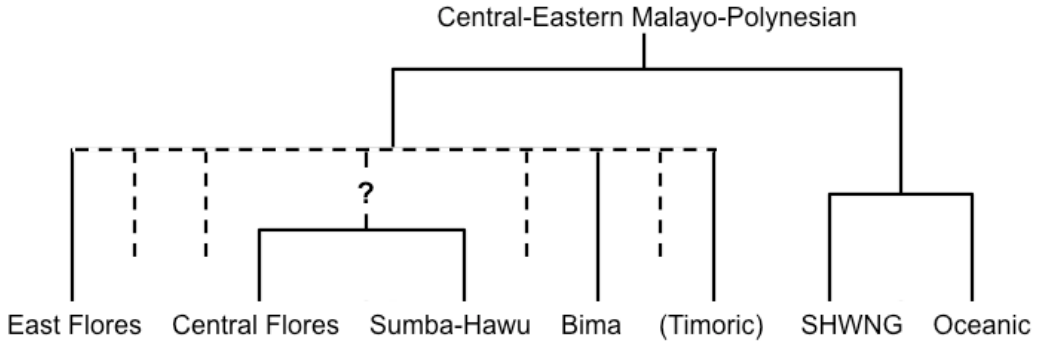
> Our trees place Bima closer to the Sumba-Hawu languages, than the neighboring languages of Flores-Lembata. Blust (2008) argues that whilst Bima-Sumba might not exist, a higher order subgrouping might exist that includes the Ambon-Timor languages along with the Sumba-Hawu and Bima languages. Our results match this interpretation.

This subsequent claim ignores Blust's clear statement (2008: 90), in the same article, that:

> In particular, it is doubtful that Esser's "Ambon-Timor" group has any more validity than "Bima-Sumba." Blust goes on to state that 'there is a Sumba-Hawu group, and perhaps a larger grouping of Sumba-Hawu with several languages of western and central Flores, but that Bimanese is excluded from this and any other group that does not also include "Ambon-Timor" languages such as Tetun.

Greenhill & Gray claim that, by clustering Bima and Sumba-Hawu together, they match Blust's 'interpretation' that 'Bimanese is excluded from this and any other group that does not also include 'Ambon-Timor' languages such as Tetun'. Tetun is present on the tree in Gray et al. (2009), and so the topology of their match can be tested against Blust's claim. It is clear that Blust (2008) considers Bima, 'Ambon-Timor', and Sumba-Hawu to all be 'Central Malayo-Polynesian' languages, and so it is reasonable to infer a tree such as the one shown in Figure S1 (the CMP languages are those in the left of the two primary branches).

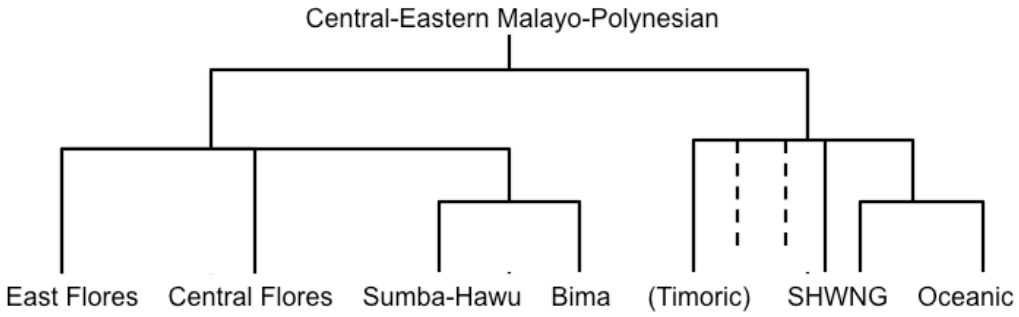
Figure S1. The relationship between Bima, Tetun, and Sumba-Hawu (following Blust 2008).



Gray et al. (2009) group the Timoric languages together with Bima and Sumba-Hawu. The Timoric languages are, however, subgrouped in an unnamed clade together with various Eastern Malayo-Polynesian languages, including Oceanic languages such as Tolai, Samoan and Hawaiian. (Gray et al. 2009) report this cluster with a posterior probability of 0.82, a level which that article elsewhere describes as showing 'strong support'. Their support for Bima, Sumba-Hawu, Central Flores and East Flores is 100%. Importantly, it is clear that the very tight grouping of Bima with Sumba-Hawu reported by Gray et al. (2009) does not 'match [Blust's] interpretation', and that the higher order subgroup that Gray et al. (2009) resort to, Central-

Eastern Malayo-Polynesian, is not of the same level as that envisaged by Blust (2008). Gray et al.'s (2009) (abbreviated) tree for the relevant languages is shown in Figure S2.

Figure S2. The relationship between Bima, Tetun, and Sumba-Hawu (following Gray et al. 2009).



Quantifying the level of mismatch between these two trees is not simple, since of the 400 languages included in Gray et al. (2009), almost half are found in the South Halmahera/West New Guinea (SHWNG) and Oceanic groups, compared to 69 which match Blust's Central Malayo-Polynesian group (between 'East Flores' and '(Timoric)' in Figure S1). Again, Greenhill & Gray's notion of providing support appears to consider only recall, and not precision.

The third of our objections that were dismissed out of hand concerns Paiwan. Greenhill & Gray write in their supplement that 'Donohue et al. mistakenly chastise us for placing Paiwan inside Malayo-Polynesian', whereas our observation of their tree is that 'Paiwan is incorrectly assigned by Gray et al. (2009) to a subgroup containing Malayo-Polynesian to the exclusion of other mainland Taiwanese languages'. The node containing Paiwan and Malayo-Polynesian has a 0.74 posterior probability, a level which Gray et al. elsewhere describe as representing a 'strongly supported' node. As we note, 'Such a subgrouping is not advocated anywhere in the literature on Austronesian languages, and is not supported by any work on the languages of Taiwan or on Austronesian historical linguistics'.

Greenhill and Gray further state that '[f]our of the other putative misplacements reflect long-standing classification difficulties or potential subgroupings (Irarutu, Kei, Maloh, Mussau)'.[1] Referring to earlier debate about the position of Irarutu (from 1961, 1978, 1989 and 1993), Greenhill and Gray state that 'rather than incorrect, our analyses are reflecting this classificatory difficulty by placing Irarutu between the Central Malayo-Polynesian and South Halmahera/West New Guinea languages'. Surprisingly they fail to cite Ross (1995), which provides convincing argumentation, based on shared sound changes, about the position of Irarutu within the South Halmahera/West New Guinea subgroup. When Greenhill and Gray state that 'our trees are more conservative' (by failing to detect the SHWNG relationship, because 'Irarutu has also undergone contact with Koiwai and other Central Malayo-Polynesian languages'), they can only be

[1] In passing we note that Greenhill and Gray refer to 'our language informant' in connection with Irarutu; we were unaware that they had conducted their own work, but are surprised that they do not cite Matsumura (1991), or Matsumura and Matsumura (1991).

referring to conservatism in terms of reliance on older references and a failure to adopt classifications based on more recent argumentation.[2]

We know of no debate on the position of Kei or Mussau that bears on our critique: noone has claimed that Kei should belong to the Yamdena-North Bomberai group to the exclusion of other languages of the region, nor that Mussau should be subgrouped with Meso-Melanesian language to the exclusion of other Oceanic languages.

Greenhill and Gray's comment on the position of Maloh is confusing. Greenhill and Gray note in their supplementary materials that '[t]he analyses here are conservative and are refusing to link Maloh more closely with other languages'. We agree that they fail to link Maloh (a Tamanic language) with the South Sulawesi languages to thus reproduce the Greater South Sulawesi subgroup. We fail, however, to see how 'conservative' can be used to describe their analyses, given the complete acceptance of this group (which is based on shared phonological and morphological innovations, and disguised by loans from Malayic languages).

'Highly congruent with the traditional subgroupings'

We note that some of the papers Greenhill & Gray cite to demonstrate the robustness of their method have already attracted critiques. For example, the discussion of an Indo-European tree that treats Romance and Germanic as a subgroup, and which provides 100% support for a separate Baltic clade within Balto-Slavic, or the unproblematic division of Uto-Aztecan into northern and southern branches (as reported in Gray and Atkinson 2003, Dunn et al. 2011; see *Linguistic Typology* 15 for responses to this second paper).

Below we address the concerns that Greenhill & Gray raise.

Concern 1. Our methodology fails to distinguish innovations from retentions

There appears to be some misunderstanding of our term 'replication', from information theory, which Greenhill & Gray assume (from their response) to be equivalent to their 'provide support for', whereas this is not (to judge from the cases in which they find support in their method) true. As we discussed in the published rejoinder, innovations are more than simply lexical, but include innovations in sound change. Given a correspondence set p:f:h;Ø, the probability techniques that Greenhill & Gray demonstrate are irrelevant to the determination that /p/ is a retention, and the other reflexes demonstrate innovative change. Further, within the f:h:Ø subset /f/ reflexes are retentions, and the non-oral reflexes are the result of innovative change.

Concern 2. Information about sound changes is completely neglected

Greenhill & Gray failed to notice that we were referring to sound changes within cognate classes; we laid out our understanding of their method quite clearly, and our concerns about the neglect of sound change was phrased as 'Examining the presence of a cognate of that particular class, not taking account of sound changes', following the expert detection of those cognates. If we had wished to impute a failure to check *regular sound correspondences* on the part of the

[2] They similarly refer to 'Remote Oceanic', for which they provide 1959 and 1983 references, ignoring work in the following decades which finds no evidence for such a grouping.

experts, we would have been more explicit. We did not discuss *regular sound correspondences*, but rather sound *change*. In the short illustrative example we provided in Table 1 we noted that the problem was neglecting a subgrouping change (*l > *r*) in the data, and Greenhill & Gray have confirmed that this kind of information, the backbone of the use of sound change in the comparative method, is entirely neglected in their method. Consequently, we are justified in our use of the quote (in the main article) from von der Gabelenz about the futility of linguistic comparison without sound correspondences. Greenhill & Gray state that:

> Modelling sound change would require a vast increase in the number of parameters to be estimated. This is a very interesting area of future research (e.g. Bouchard-Côté et al. 2009), but such enormously complex models are not required to accurately infer linguistic relationships.

We believe that *past* research has shown that sound change is a very interesting area of research, and work such as Nakleh at al. (2005) demonstrates that it is not beyond computational implementation.

Concern 3. Lexical items are frequently borrowed and hence 'Gray et al.'s method is as likely to detect the cumulative effects of lexical borrowing … as it is to detect historical developments resulting from original differentiation from a proto-language'

Greenhill & Gray note, in the supplementary materials, numerous cases where the misclassification of language subgroups is due to 'undetected borrowing', and this is entirely our point: borrowing plays a massive role in the distribution of lexical items and ultimately undermines their phylogentic interpretation. In their response, misclassifications are accounted for using the following argumentation (only a selection is presented): (different clustering) 'due to unidentified lexical borrowings'; 'may reflect unidentified borrowings between these two neighboring subgroups'; 'may reflect the widespread diffusion of features'; 'borrowings are the likely explanation for the minor mismatch between our results and the traditional linguistic subgroupings'; and so on.

Greenhill & Gray make much of low rates of borrowing for the kinds of lexical items that they include in their wordlists, but when they state in their supplement that 'We have been reassessing the cognate coding in that area, and have uncovered 17 previously unrecognized loan words in the Sama-Bajaw language Inabaknon' (taking the total identified to 32, 15%), we have to wonder how low the rates are. Citing work published elsewhere (Nelson-Sathi et al. 2011), they claim an average borrowing rate in, for instance, Indo-European languages of 8%. They are perhaps misled into this belief by relying on the annotated wordlists by Kruskal, Conrad and Black (1997), which are focussed on identifying cognate classes rather than loans.

It is quite possible for a language to borrow from another language within the expected cognate class – witness English *shirt* (inherited) and *skirt* (borrowed from Norse). Examining a 200 item Swadesh list for English one finds 16 loans from Norse, 13 from French, 2 from Low

German and 1 from Dutch, for a total 16% of the wordlist, double the 'average value' Greenhill & Gray cite.[3]

We know of no reason to believe that English is exceptional in this regard, and again cite Grant's (2005) work on Chamic languages. Greenhill & Gray correctly point out that Jarai is not included in their trees (though it is found in their database), and note that '[t] he critical issue is not whether there are high levels of borrowing, but whether there are high levels of borrowing in our sample of Austronesian basic vocabulary'. We note the case of Hainan Cham (Utsat), closely related to Jarai, which Greenhill & Gray list in their database as having 24 loans in a wordlist with 210 senses. We find that their own notes (accessed March 12, 2012) identify an additional 24 loans, for a total 22.9% of the items listed presenting loans. This figure is less than the 48% loan rate that can be deduced for Jarai from Grant's data, but higher than the 1.5% of items identified as loans in Jarai in Greenhill & Gray's database. Similarly Phan Rang Cham is listed as having 10 loanwords in their 210 item list, while their own notes identify an additional 11 items, with annotations such as 'borrowed from Mon-Khmer'.

The reader will note a consistent pattern of underestimating the number of loans by a factor of approximately two (in languages contended by Greenhill & Gray). This higher loanword rate represents a more realistic estimate of the number of loans in the languages examined. Do we have any reason to believe otherwise, given the admission of frequent instances of 'undetected borrowing'? We note that other misplaced languages are excused by an appeal to lexical diffusion, which is precisely the point we make in our original article. We quote the following examples from Greenhill & Gray's supplementary materials:

- 'due to unidentified lexical borrowings … does not invalidate our support for the North New Guinea subgroup'
- 'Irarutu has also undergone contact with Koiwai and other Central Malayo-Polynesian languages from the Bomberai peninsula. This contact may be the reason why our trees … do not weakly subgroup Irarutu with the South-Halmahera/West New Guinea languages'
- 'may reflect unidentified borrowings between these two neighboring subgroups'
- 'may reflect the widespread diffusion of features'
- 'our results may in fact be confirming yet to be published results'
- 'borrowings are the likely explanation for the minor mismatch between our results and the traditional linguistic subgroupings'
- 'The slight misplacement of a single language does not mean that Meso-Melanesian is not supported by our results'

[3] The words are: split (Dutch); animal, because, count, dig, flower, fruit, lake, mountain, person, push, river, turn, vomit (French); pull, rub (Low German); (tree) bark, big, cut, die, dirty, egg, give, hit, husband, leg, root, rotten, skin, sky, they, wing (Norse). The entire list is reproduced at the end of these materials.

- 'may either reflect contact-induced change with neighboring Sulawesi languages, although there is little evidence for contact'
- 'these differences could be explained by unidentified borrowings between languages within these subgroups'

Gray et al. (2009) frequently refer to 'slight misplacement', but are not willing to acknowledge that a mistake is a mistake. They have raised the example of Albanian previously being mis-assigned to Indo-Iranian. To bring the matter home, if English were assigned to Indo-Iranian would we be happy to treat the resulting tree as correct with some languages showing a 'slight misplacement'?

Concern 4. There are clear discrepancies in the placements of individual languages between the phylogeny and the expected language relationships

Let us clarify our position: Given that Greenhill & Gray's lexical tree of the AN-language family is based on lexical cognates identified by historical linguists who are experts in the family, Greenhill & Gray's tree should be expected to follow the phylogenetic tree closely. Any unexpected departure from that tree needs to be examined critically, since their tree was derived using data which are not independent of the comparative method. While we are interested in examining any misplacement of individual languages and whole subgroups, we are particularly concerned that these misplacements are not random. Rather, whenever these misplacements are found Greenhill & Gray's method picks up (social) geography in preference to subgrouping derived from phylogenetic inheritance. In *all* cases where such a discrepancy arises we have shown that Gray et al.'s (2009) method replicates geography rather than linguistic subgrouping. Rather than critiquing the placement of 37 languages (as Greenhill & Gray characterise our article), we critique the claim (Gray et al. 2009: 479) that '[t]he major features of our trees are congruent with the results of the comparative method'. The importance of such congruence evaporates when examined critically (as discussed above with respect to replication).

Concern 5. Polynesian subgroupings are completely scrambled

Greenhill & Gray note that rather than being fully resolved, the internal subgrouping of Polynesian is an 'active area of debate', and we acknowledge that this is not just true, but almost inevitable given the extensive contact and borrowing that has occurred between the languages concerned. Nevertheless, compared to the lack of consensus found in any other Austronesian subgroups of similar scope, Polynesian is a model of relative clarity.

Gray et al. (2009) classify Polynesian without replicating the two first-order divisions, as discussed in our paper. Greenhill & Gray show us that alternative algorithms using their data similarly fail to replicate the two-way split between Tongic (Tongan and Niuean, so closely related that they are almost dialects of the one language), on the one hand, and all other Polynesian languages, on the other hand. *All* historical linguists concerned with Polynesian subgrouping accept this fundamental division (eg., Marck 2000 and all others). Consequently, Greenhill & Gray err in asserting that 'the only Polynesian subgroups that are relatively uncontroversial' 'are replicated in our trees', because from the outset, their internal classification of Polynesian fails. Discussing their whimsical grouping of Tongan with Samoan, Greenhill &

Gray assert that neither ongoing borrowing between these languages nor conservativism are exceptional for this pair, and so cannot be named as responsible for the misclassification. Regardless of the cause, Gray et al.'s method results in a classification that is refuted by the evidence of sound correspondences.

Concern 6. The degree of similarity between our results and the comparative method are overstated

Greenhill & Gray note that the unrooted tree we provided for illustrative purposes replicates the subgrouping of the European languages it contains with accuracy of 84%, but state that 'Despite this putatively high accuracy, the tree has some notable misplacements.' This is exactly our purpose in using this tree as an illustration: 84% is not a high level of replication, and 81% (the degree to which Marck's Polynesian tree is replicated) is even less.[4] The point we make with this tree, which was produced with Splitstree (Huson and Bryant 2006), but could for our purposes just as easily have been invented, is that an 84% replication rate might (to use the language used in Gray et al. article and in Greenhill & Gray's reply) provide 'strong support', but it is no guarantee of accuracy. We note that using completely unrelated data (the typological features coded in WALS) we achieved the same level of inaccuracy in replicating European subgroups that Gray et al. (2009) did with Polynesian, when they were using the fruits of the comparative method, in the sense that their data set consisted of lexemes that had been coded for cognacy (after taking into account regular sound change). Their tree, then, is an illustration of the dangers of not paying attention to the direction of sound change.

Concern 7. The trees represent 'distance-decay and local borrowing' rather than phylogeny

Greenhill & Gray state that geography does not play a significant role in the distribution of words: 'If the trees reflected geography rather than genealogy then … Maori would group with the languages of New Caledonia rather than Eastern Polynesian languages … .' We agree that it is indeed not the case, but point out that their analogy is a false one. As we were at pains to emphasise in our article, we were not speaking of Euclidean geometry, but of 'geography or social distance', later rephrased in the same article as 'human geography (that is, social distance)', 'social space (typically geography but also possibly social networks that bypass immediate neighbours)', and 'geography/social contact'. It is surprising that Greenhill & Gray did not pick up on the 'social geography' sense that we repeatedly emphasised.

Rather than '[p]ointing out minor misplacements of individual languages' some examples to discuss in our supplementary materials, as Greenhill & Gray assert, we examined '*all* cases where we find a mismatch between geography and phylogeny' (our abstract, with emphasis added). We later repeat this statement: '*All* discrepancies between phylogenies represent either a geographically distant language failing to be grouped with its subgroup cousins, or else a proximal language from a separate subgroup being falsely clustered with its geographic neighbors' (emphasis original). To restate: '*all* cases of successful reproduction are in

[4] The method used to quantify the accuracy of the tree is described in Donohue et al. (2011), drawing on information theory (van Rijsbergen 1979).

geographically uncomplicated regions … the subgroups which are least well replicated … are *all* in socially complex regions'.

General point

Greenhill & Gray upbraid us for the perception that we 'treat the consensus tree we reported as a single tree rather than as a visual summary of the posterior distribution of trees'. In fact, as we noted,

> The tree presented by Gray et al. (2009) is a 'consensus tree', representing a 'best-fit' compromise between a number of possible trees. In the interests of having a testable hypothesis to evaluate, and since this tree has been presented as showing congruency, we shall evaluate it as a claim.

If the tree is not intended to be evaluated, we are not sure why it was published. If no subgrouping hypothesis is meant to be evaluated (as is implied by the presentation of Greenhill & Gray's Figure S1), then the method is not one that can be taken seriously.

Conclusion

We are excited at the idea that computational tools can offer first approximations of new directions for historical linguists to investigate – but emphasis that these can only be first approximations, unless they fully apply the comparative method (see, for instance, Brown et al. 2011).

Additional references

Brown, Cecil H., David Beck, Graegorz Kondrak, James K. Watters & Søren Wichmann. 2011. "Totozoquean". *International Journal of American Linguistics* 77:3.323–372.

Donohue, Mark, Simon Musgrave, Bronwen Whiting & Søren Wichmann. 2011. "Typological Feature Analysis Models Linguistic Geography". *Language* 87:2.369–383.

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson & Russell D. Gray. 2011. "Evolved Structure of Languages Shows Lineage-specific Trends in Word-order Universals". *Nature* 473.79–82.

Matsumura, Takashi. 1991. "Irarutu Phonology". *Workpapers in Indonesian Languages and Cultures* 10.37–74.

Matsumura, Takashi & Michiko Matsumura. 1991. "A Preliminary Grammar Sketch of the Irarutu Language". *Workpapers in Indonesian Languages and Cultures* 10.75–110.

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

Swadesh 200-item list

The Swadesh list we used to calculate loans in modern English is shown below, with the provenance of different lexemes indicated.

	Item	Provenance
1	all	Proto-Germanic
2	and	Proto-Germanic
3	animal	**French**

4	ashes	Proto-Germanic
5	at	Proto Indo-European
6	back	Proto-Germanic
7	bad	Old English?
8	bark (of a tree)	**Norse**
9	because	**French**
10	belly	Proto-Germanic
11	big	**Norse**
12	bird	Old English
13	to bite	Proto Indo-European
14	black	Proto-Germanic
15	blood	Proto-Germanic
16	to blow (wind)	Proto Indo-European
17	bone	Proto-Germanic
18	to breathe	Proto-Germanic
19	to burn (intrans)	Proto-Germanic
20	child (young)	Old English
21	cloud	Proto-Germanic
22	cold (weather)	Proto Indo-European
23	to come	Proto Indo-European
24	to count	**French**
25	to cut	**Norse**
26	day (not night)	Proto-Germanic
27	to die	**Norse**
28	to dig	**French**
29	dirty	**Norse**
30	dog	Old English
31	to drink	Proto-Germanic
32	dry (substance)	Proto-Germanic
33	dull (knife)	Proto Western Germanic
34	dust	Proto-Germanic
35	ear	Proto Indo-European
36	earth (soil)	Proto-Germanic
37	to eat	Proto Indo-European
38	egg	**Norse**
39	eye	Proto Indo-European
40	to fall (drop)	Proto Indo-European
41	far	Proto Indo-European
42	fat (substance)	Proto-Germanic
43	father	Proto Indo-European
44	to fear	Proto-Germanic
45	feather (large)	Proto-Germanic
46	few	Proto Indo-European

47	to fight	Proto Indo-European
48	fire	Proto Indo-European
49	fish	Proto Indo-European
50	five	Proto Indo-European
51	to float	Proto-Germanic
52	to flow	Proto-Germanic
53	flower	**French**
54	to fly	Proto Indo-European
55	fog	ModernEnglish
56	foot	Proto Indo-European
57	four	Proto Indo-European
58	to freeze	Proto Indo-European
59	fruit	**French**
60	to give	**Norse**
61	good	Proto-Germanic
62	grass	Proto-Germanic
63	green	Proto-Germanic
64	guts	Proto-Germanic
65	hair	Proto-Germanic
66	hand	Proto Indo-European
67	he	Proto Indo-European
68	head	Proto Indo-European
69	to hear	Proto Indo-European
70	heart	Proto Indo-European
71	heavy	Proto-Germanic
72	here	Proto-Germanic
73	to hit	**Norse**
74	hold (in hand)	Proto-Germanic
75	how	Proto Indo-European
76	to hunt (game)	Proto-Germanic
77	husband	**Norse**
78	i	Proto Indo-European
79	ice	Proto-Germanic
80	if	Proto-Germanic
81	in	Proto Indo-European
82	to kill	Old English
83	know (facts)	Proto Indo-European
84	lake	**French**
85	to laugh	Proto-Germanic
86	leaf	Proto-Germanic
87	left (hand)	Proto Western Germanic
88	leg	**Norse**
89	to lie (on side)	Proto Indo-European

90	to live	Proto-Germanic
91	liver	Proto-Germanic
92	long	Proto Indo-European
93	louse	Proto Indo-European
94	man (male)	Proto Indo-European
95	many	Proto Indo-European
96	meat (flesh)	Proto-Germanic
96b	moon	Proto Indo-European
97	mother	Proto Indo-European
98	mountain	**French**
99	mouth	Proto Indo-European
100	name	Proto Indo-European
101	narrow	Proto-Germanic
102	near	Proto-Germanic
103	neck	Proto-Germanic
104	new	Proto Indo-European
105	night	Proto Indo-European
106	nose	Proto Indo-European
107	not	Proto Indo-European
108	old	Proto-Germanic
109	one	Proto Indo-European
110	other	Proto Indo-European
111	person	**French**
112	to play	Proto Western Germanic
113	to pull	**Low German**
114	to push	**French**
115	to rain	Proto-Germanic
116	red	Proto Indo-European
117	right (correct)	Proto Indo-European
118	right (hand)	Proto Indo-European
119	river	**French**
120	road	Proto-Germanic
121	root	**Norse**
122	rope	Proto-Germanic
123	rotten (log)	**Norse**
124	rub	**Low German**
125	salt	Proto Indo-European
126	sand	Proto Indo-European
127	to say	Proto Indo-European
128	scratch (itch)	MiddleEnglish
129	sea (ocean)	Proto-Germanic
130	to see	Proto Indo-European
131	seed	Proto Indo-European

132	to sew	Proto Indo-European
133	sharp (knife)	Proto-Germanic
134	short	Proto Western Germanic
135	to sing	Proto-Germanic
136	to sit	Proto Indo-European
137	skin (of person)	**Norse**
138	sky	**Norse**
139	to sleep	Proto Indo-European
140	small	Proto Indo-European
141	to smell (perceive odor)	Old English
142	smoke	Proto-Germanic
143	smooth	Proto Indo-European
144	snake	Proto-Germanic
145	snow	Proto Indo-European
146	some	Proto Indo-European
147	to spit	Proto-Germanic
148	to split	**Dutch**
149	to squeeze	ModernEnglish
150	to stab (or stick)	MiddleEnglish
151	to stand	Proto Indo-European
152	star	Proto Indo-European
153	stick (of wood)	Proto-Germanic/**Norse**
154	stone	Proto-Germanic
155	straight	Proto-Germanic
156	to suck	Proto Indo-European?
157	sun	Proto Indo-European
158	to swell	Proto-Germanic
159	to swim	Proto-Germanic
160	tail	Proto-Germanic
161	that	Proto Indo-European
162	there	Proto-Germanic
163	they	**Norse**
164	thick	Proto Indo-European
165	thin	Proto Indo-European
166	to think	Proto-Germanic
167	this	Proto Indo-European
168	thou/you	Proto Indo-European
169	three	Proto Indo-European
170	to throw	Proto Indo-European
171	to tie	Proto-Germanic
172	tongue	Proto Indo-European
173	tooth (front)	Proto Indo-European
174	tree	Proto-Germanic

175	to turn (veer)	**French**
176	two	Proto Indo-European
177	to vomit	**French**
178	to walk	Proto Western Germanic
179	warm (weather)	Proto Indo-European
180	to wash	Proto-Germanic
181	water	Proto Indo-European
182	we	Proto Indo-European
183	wet	Proto Indo-European?
184	what	Proto Indo-European
185	when	Proto Indo-European
186	where	Proto Indo-European
187	white	Proto Indo-European
188	who	Proto Indo-European
189	wide	Proto-Germanic
190	wife	Proto-Germanic
191	wind (breeze)	Proto Indo-European
192	wing	**Norse**
193	wipe	Proto-Germanic
194	with (accompanying)	Proto Indo-European
195	woman	Old English
196	woods	Proto Indo-European
197	worm	Proto Indo-European
198	ye	Proto Indo-European
199	year	Proto Indo-European
200	yellow	Proto Indo-European

Additional discussion of misclassified languages

Nakanai and the North New Guinea languages

Greenhill & Gray suggest that the failure to include Nakanai with the Meso-Melanesian languages 'was presumably due to unidentified lexical borrowings between these Willaumez languages and the neighboring languages of West New Britain belonging to the Meso-Melanesian subgroup'. We agree.

Meso-Melanesian

Greenhill & Gray state that we 'claim that our [Greenhill & Gray's – DOO] placement of Mussau with the Meso-Melanesian subgroup invalidates this grouping'. We did not use the word 'invalidate'; we stated that Mussau should not be included in that grouping 'for accurate replication' (supplementary materials). Stating that 'the placement of Vitu at the base of this subgroup is not particularly surprising given that the Bali-Vitu lineage is thought to be a primary branch of Meso-Melanesian (Lynch et al. 2002)'. obscures the fact that Vitu is placed further

from the rest of the Meso-Melanesian languages than Mussau, which is not a Meso-Melanesian language.

South Halmahera/West New Guinea and Eastern Malayo-Polynesian
These subgroups (SHWNG is a daughter of EMP) should include the language Irarutu. Since linguistic evidence firmly subgroups Irarutu with South Halmahera/West New Guinea (Ross 1995), the exclusion of Irarutu from the rest of the West New Guinea languages (which are located in a north-facing bay; Irarutu is in a south-facing bay) reflects the contact that Voorhoeve (1989) observed between Irarutu and Koiwai and other Central Malayo-Polynesian languages. Greenhill & Gray note that '[t]his contact may be the reason why our trees are more conservative and do not weakly subgroup Irarutu with the South-Halmahera/West New Guinea languages', though we do not understand how this is 'conservative'.

Central Maluku
Greenhill & Gray's tree includes the languages of Aru with those of Central Maluku. As Greenhill & Gray note, '[t] he placement of the two languages Ujir and Ngaibor from Aru in a group together with 13 Central Maluku languages may reflect unidentified borrowings between these two neighboring subgroups in Maluku'. We agree; contact is the likely explanation.

Yamdena-North Bomberai
As Greenhill & Gray note, Koiwai has not been grouped with the Yamdena-North Bomberai group proposed by Blust (1993). Most likely, as Greenhill & Gray observe, is a contact explanation: 'the placement of Koiwai here may reflect the widespread diffusion of features such as glide truncation across the Bomberai region'. Concerning the inclusion of Kei in this cluster, Greenhill & Gray note that '[i]nstead of being incorrect, our results may in fact be confirming yet to be published results'. We welcome the publishing of these results, and point out, in our original article that 'claims of accurate replication must rely on the replication of established groups and cannot claim that the generation of speculative new proposals is proof of matching results'. No established group of Austronesian languages included Kei with the North Bomberai languages.

Greater South Sulawesi
Greenhill & Gray's version of Greater South Sulawesi does not include the Tamanic languages. Since Tamanic and South Sulawesi are the two first-order subgroups that constitute Greater South Sulawesi, failure to group the two sub-families together cannot be reconciled with Greenhill & Gray's statement that 'this does not indicate a lack of support for the Greater South Sulawesi language subgroup'. It is not clear what they mean when they state that '[t] he analyses here are conservative and are refusing to link Maloh [the Tamanic language in their database – DOO] more closely with other languages'.

Figure S3. The Greater South Sulawesi group (Blust 2009).

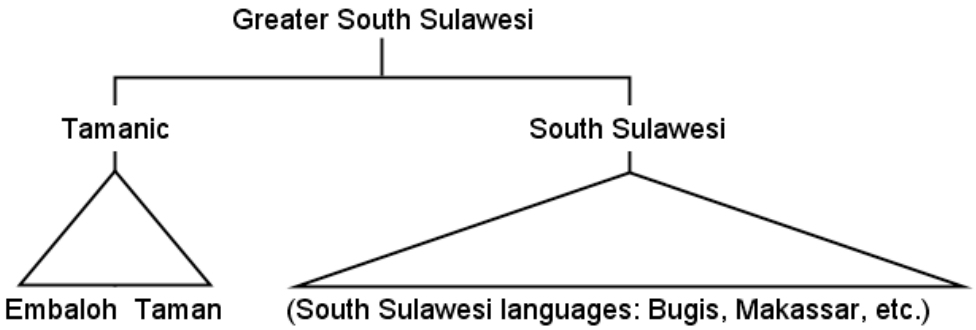
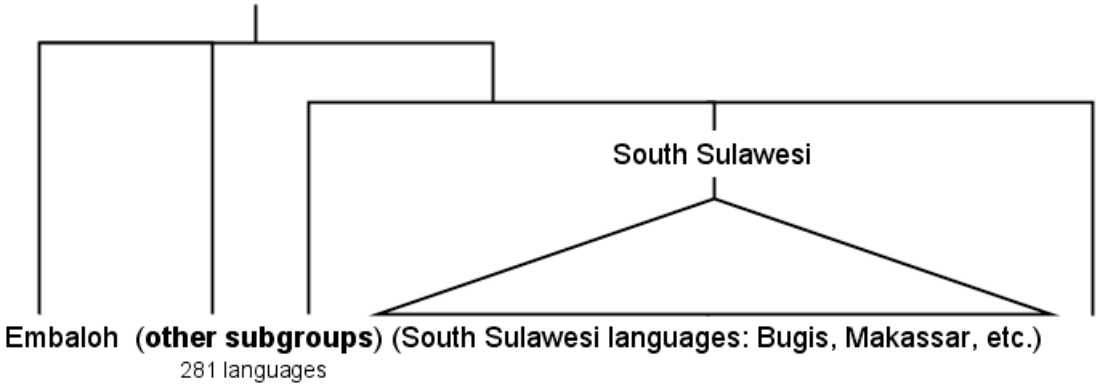


Figure S4. The Greater South Sulawesi group (Gray et al. 2009).



Barito

Greenhill & Gray's tree groups the Barito languages with the North Borneo subgroup, and not with Sama-Bajaw, their sister (Blust 2010). Greenhill & Gray state that they 'have been reassessing the cognate coding in that area, and have uncovered 17 previously unrecognised loan words in the Sama-Bajaw language Inabaknon'. Since '[t]hese borrowings are the likely explanation for the minor mismatch [= complete misplacement - DOO] between our results and the traditional linguistic subgroupings in this region'. We agree.

Sangiric

As Greenhill & Gray note, 'The Sangiric languages are located in the Sulawesi region but are definitely Philippines languages', yet they are not placed in the Philippine group in Greenhill & gray's tree. Greenhill & Gray suggest that 'Our placement of Sangiric as a deeper group within the Western Malayo-Polynesian linkage may either reflect contact-induced change with neighboring Sulawesi languages, although there is little evidence for contact'. While the Philippines group is elsewhere clustered, contact has removed the Sangiric languages, thus preventing successful replication.