New methodologies for historical linguistics?
Calibrating a lexicon-based methodology for diffusion vs. subgrouping*

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

Recent research claims that analysis of lexical cognate classes for a basic wordlist can reproduce linguistic subgroups within the Austronesian family (Gray et al. 2009). The analysis is open to question in two respects. Primarily, the lexically-based classification, primed with pre-established cognate classes of the family it seeks to emulate, fails to differentiate shared retentions from shared innovations. Secondly, languages and language families typically disperse through contiguous regions (especially in the Pacific) which means that geography or social distance should be expected crudely to match phylogeny in most cases. The reproduction fails because of local borrowing between branches not closely related to each other. For instance, when we examine disjunct distributions, cases in which the phylogeny does not match a straightforward geographic spread, we can determine which of these (phylogeny or geography) the lexical cognate approach preferentially detects. Where we find a mismatch between geography and phylogeny, Gray et al.’s approach clusters languages based on human geography (that is, social distance), not linguistic subgroup. In all cases of divergence between Gray et al.’s tree and accepted Austronesian trees, the discrepancy is a product of the former representing social distance rather than historical phylogenetic relationships. In summary, the examination of lexical cognate classes is not a valid proxy for the comparative method, though it is a useful heuristic for detecting pairs of languages that are either lexically conservative, or which show the effects of later lexical diffusion (without discriminating between these two outcomes).

Keywords: computational methods, historical linguistics, Polynesian, Austronesian, areality, lexical borrowing

* Many thanks are due to the constructive comments of three referees and the journal editors; they have greatly improved this paper. Alex François enabled our engagement with French.
1. Introduction

Despite there being a number of established, and proven, methodologies for work in historical linguistics, a number of alternative techniques have been popularised. We welcome new techniques that can help unravel linguistic histories, but attention is needed to evaluate what these techniques actually measure. To that end, we critique and evaluate one such lexically-based study, Gray et al. (2009). Elements of Gray et al’s study are representative of similar works (e.g., Dunn et al. 2008), and our rubric of evaluation should serve as a test for them as well (e.g. Donohue 2012; see Pompei et al. 2011 for alternative evaluation metrics).

Gray et al. (2009, cf. Gray & Jordan 2000 and Gray & Atkinson 2003) argue that a classification of languages based on the identities of lexical cognate classes found in basic vocabulary lists (employing Bayesian statistical techniques), detects linguistic ancestry and degree of relatedness among related languages. The methodology (described in detail in Gray et al. 2009, Supporting Online Materials) essentially involves:

- Examining lexical correspondences in a (210-item) basic word list;
- Dividing the items of the basic wordlist into cognate classes, already determined by experts in Austronesian historical linguistics;
- Examining the presence of a cognate of that particular class, not taking account of sound changes;
- Ignoring the differences between similarities that reflect shared retentions and those that reflect shared innovations; and,
- Ignoring phonological or morphological material.

The method can be illustrated in miniature with an Austronesian dataset based on Greenhill et al. (2008: 276). An example of the kinds of data and the way they have been coded are shown in Table 1. Here four items of basic vocabulary are arranged by language (all members of the Oceanic subgroup of Austronesian, and all but Motu members of the Polynesian subgroup of Oceanic). The lexemes are sorted into clusters depending on the cognacy of the terms compared in each column. Thus reflexes of *lima “hand” in the first column do not provide any information to distinguish different clusters of languages in this mini dataset; on the other hand *kili “skin” in Niuean is cognate with *tiiri “skin” in Tahitian, and so they are classed together in the same cognate class (along with most of the other entries in this column). Each lexical item examined (other than “hand” in the material here) divides into a number of cognate classes (40 cognate classes for a single lexeme are not uncommon in the Austronesian Basic Vocabulary Database, Greenhill et al. 2008), and on the basis of these similarities languages are subgrouped via algorithms that (to gloss over numerous complexities) simultaneously evaluate the different
New methodologies for historical linguistics?

subgrouping signals from the different cognate clusterings arising in different lexical items (see Greenhill et al. 2008 or Gray et al. 2009 for a fuller description).

As noted, information about sound changes is not included in the technique. That “hand” and “skin” (in Tahitian, Rapa Nui and Rurutuan) both show evidence for a *l > r change, evidence for subgrouping the Eastern Polynesian languages together, is not included in the assessment of subgroups (similarly for the *k > ? change in reflexes of *kuli “skin”). Since the cognate classes are constant, the forms are treated as identical (only the innovative words in Samoan and Motu provide contrastive data for the “skin” entry). For the lexeme “plant” we find two widespread cognate classes (the innovative Rapa Nui form does not provide subgrouping information). These are not equal, however; *too is a Proto-Polynesian innovation and should not be used as subgrouping evidence for the position of Niuean, Samoan and Anuta within Polynesian. At the same time, the fact that tanu, kanu and nanu reflect Proto-Oceanic *tanu “plant” (itself an irregular reflex of Proto-Austronesian *tanam), as evidenced by the witness of the non-Polynesian Oceanic language Motu, shows that this cognate should not be taken as providing subgrouping evidence for Kapingamarangi, Hawaiian, Tahitian, Rurutuan and Marquesan, since *tanu in these languages is a conservative retention, rather than an innovation. In a similar vein, the fact that Niuean, Samoan, and Anuta anu, and Rapa Nui aanu are all reflexes of Proto-Polynesian *qanu(-si) “spit” (itself a reflex of Proto-Oceanic *qanusi)1 and this shared cognacy is counted as subgrouping evidence for these four languages even though it is a retention (compare to the innovative form for the same lexeme in Tahitian, Rurutuan and Marquesan).2

Table 1. Example of (partial cognate class organisation).

<table>
<thead>
<tr>
<th>Language</th>
<th>“hand”</th>
<th>“skin”</th>
<th>“plant”</th>
<th>“spit”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Niuean</td>
<td>lima</td>
<td>1</td>
<td>kili</td>
<td>1</td>
</tr>
<tr>
<td>Samoan</td>
<td>lima</td>
<td>1</td>
<td>paʔu</td>
<td>2</td>
</tr>
<tr>
<td>Anuta</td>
<td>lima</td>
<td>1</td>
<td>kili</td>
<td>1</td>
</tr>
<tr>
<td>Kapingamarangi</td>
<td>lima</td>
<td>1</td>
<td>gili</td>
<td>1</td>
</tr>
<tr>
<td>Hawaiian</td>
<td>lima</td>
<td>1</td>
<td>ʔili</td>
<td>1</td>
</tr>
<tr>
<td>Tahitian</td>
<td>rima</td>
<td>1</td>
<td>ʔiri</td>
<td>1</td>
</tr>
</tbody>
</table>

1. Forms such as Kowiai naʔanis “spit”, Ambai kanu and Wandamen kanisu “spit, sneeze” in Austronesian languages west of and outside the Oceanic subgroup highlight the ephemeral value of posited lexical innovations in regions where most languages are imperfectly or undocumented.

2. The comparisons, unlike the application of the comparative method, simply involve testing for identity of cognate classes given for items in a wordlist. Semantic shifts are not traced.
Table 1.

<table>
<thead>
<tr>
<th>Language</th>
<th>“hand”</th>
<th>“skin”</th>
<th>“plant”</th>
<th>“spit”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rapa Nui</td>
<td>rima</td>
<td>1</td>
<td>kiri</td>
<td>1</td>
</tr>
<tr>
<td>Rurutuan</td>
<td>rima</td>
<td>1</td>
<td>?iri</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marquesan</td>
<td>rima</td>
<td>1</td>
<td>ki?i</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Motu</td>
<td>ima</td>
<td>1</td>
<td>kopi</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The application of this method to a large body of Austronesian lexical data produces a set of subgrouping results that show a (much) greater-than-chance congruence with the subgroups of the Austronesian family constructed using the comparative method (e.g., Lynch, Ross & Crowley 2002, Pawley 2007, Blust 2009; we follow Gray et al. in taking Blust 2009 (in press as they wrote) as the benchmark for successful replication). We argue, however, that due to the fact that (prior to the age of European exploration) language families have tended to elaborate and spread across contiguous regions rather than (in the main) ‘leap-frogging’, alternative hypotheses must be examined to determine whether their method successfully replicates language families and their subgroups directly, or through the detection of a common factor shared by both lexical items and linguistic subgroups. Due to their reliance on purely lexical materials, Gray et al.’s (2009) analysis produces clusters that reflect ‘distance-decay’ and local borrowing effects more than established linguistic subgroups. Rather than geographical effects being just a “conflicting signal caused by borrowing between neighboring languages” (Gray et al. 2009:480), they are frequent enough to undermine this lexical approach as a means to replicate families and subgroups.

Lexical items are frequently borrowed between languages, and when taken as a cumulative effect through history represent the effects of social dominance and geographical proximity. Borrowing is mediated by social and historical processes. Detailed documentation of these processes is often not available, and so we see geographical location as a rough proxy for the more elusive social factors. Lexical items are widely recognised to be the elements of a language most prone to diffusion, and the least reliable (in the absence of regular sound correspondences) to determine phylogenetic relationships (e.g., Séguy 1971, Goebel 1984, Cavalli-Sforza & Wang 1986, Thomason & Kaufman 1988, Kessler 2001, McMahon & McMahon 2007).

Notes:
3. In the absence of a detailed study, we note the widespread, and successful, use of terms such as ‘Northern X’ and ‘West X’ in designating subgroups. The major subgroups of the Afro-Asiatic, Austro-Asiatic, Chibchan, Dravidian, Indo-European, Nakh-Daghestanian, Niger-Congo, Oto-Manguean, Salishan, Sepik, Sino-Tibetan, Tai-Kadai, Uralic, Uto-Aztecan and Wakashan families all show significant levels of predictability on the basis of spatial proximity. For example, Uralic languages east of 70 degrees east can be predicted to be in the Samoyedic subgroup.
New methodologies for historical linguistics? 509

2005, Meillet 1908, though see the results of Holman et al. 2008a, b with respect to a more selective wordlist). A priori, Gray et al.’s method is as likely to detect the cumulative historical effects of lexical borrowing between neighbouring groups, a largely geographical process (based on the references above on the correlation between distance and lexical similarity), as it is to detect historical developments resulting from original differentiation of a proto-language.

An analysis that effectively detects social space (typically geography but also possibly social networks that bypass immediate neighbours) can, in many cases, appear to replicate the subgroups of a linguistic family. In these cases Gray et al.’s method is not sufficiently sensitive to distinguish linguistic subgroups resulting from original language dispersal and lexical clustering resulting from continued culture contact across social space. As a result, Gray et al. are able to address “questions about human prehistory” (2009: 479), but the answers, we argue, have less to do with initial colonisation and more to do with subsequent social connectivity through (pre)history.

We present a simple heuristic that distinguishes the social/geographic signal (borrowing) from the phylogenetic signal (inheritance) in Gray et al.’s (2009) comparisons with Blust (2009), illustrated more exhaustively in our supplementary materials. In the following section we examine Gray et al.’s level of success in replicating the received tree (from the application of the comparative method) for the Polynesian languages, a subgroup for which Gray et al. (2009) report very high probabilities of successful replication. In all cases where a disparity arises, geographical proximity rather than subgrouping most parsimoniously explains the clusters identified by Gray et al. (2009). Further, it is readily apparent that two (or more) lexically conservative languages are frequently falsely grouped together (see Campbell & Poser 2008).

2. Linguistic evidence for phylogenies

Language families most commonly disperse through contiguous regions, and can remain contiguous following dispersal. For this reason a model that claims to reproduce linguistic subgrouping can only be proven not to reflect social space (the socially mediated network of spatial interactions) if it performs satisfactorily in linking geographically noncontiguous subgroups of a family and in separating geographically contiguous, but genealogically distant, subgroups. This is reflected in Pawley’s dictum on rigour in establishing reconstructions (referring to the

4. The online supplementary materials offer discussion of the other ways in which major subgroups described in Blust (2009) are not successfully replicated by Gray et al. (2009).
Austronesian family, but more widely applicable): “the conditions should be such that borrowing is an unlikely explanation for the distribution” (Pawley & Pawley 1994: 334). This reflects Meillet’s concerns about the putative value of lexical items: “Les coincidences de vocabulaire n’ont en general qu’une très petite valeur probante” (1908: 126, also 17–23, 130).

Gray et al. (2009) claim to “support 26 of 34 putative Austronesian language subgroups and linkages discussed [in Blust (2009)]” (76%); the derived tree is said to be “congruent with the traditional subgroups identified by phonological and morphological evidence” (2009:479). But as acknowledged in their Supporting Online Materials, many of their clusters (putative subgroups) and hierarchical strata are unsupported by linguistic opinion, and many accepted linguistic subgroups are poorly represented or show divergent internal structures in their results. For instance, Gray et al. make comments such as “Two microgroups, however, are not included with these but are located with their regional neighbors in Sulawesi” (2009, Supporting Online Materials: 9, referring to the southernmost members of the Greater Philippines subgroup), and “Our results group the North New Guinea and Papuan Tip families together as New Guinean Oceanic, but do not include the Meso-Melanesian languages. Instead, Meso-Melanesian is grouped with the St. Matthias language Mussau and the South-East Solomonic languages” (2009, Supporting Online Materials: 7; the range of Meso-Melanesian borders on that of the South-East Solomonic languages in the south and of Mussau in the north, while North New Guinea and Papuan Tip are neighbours to the west). While new hypotheses are valuable, these cannot be counted as ‘replication’ of traditionally-accepted nodes. To summarise the material that follows, only 15 of the 31 (48%) subgroups listed in Blust (2009) between Malayo-Polynesian and Polynesian are fully and accurately replicated by the clustering analysis (compare Figures 1A and 1B). 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 (either as a result of shared lexical retentions, or later lexical diffusion).

A topological comparison of the trees for Austronesian shows clear discrepancies between Blust (2009) and Gray et al. (2009) (Figure 1). A reconstruction of the tree based on Blust (2009) is much flatter and less hierarchical than in Gray et al. (2009). Some nodes generated by Gray et al. have been explicitly rejected by historical linguists (e.g., the treatment of Bima-Sumba in Blust 2008). Numerous other nodes and subgroups are completely unsupported by non-lexical linguistic

---

5. 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, and since this tree has been presented as showing congruency, we shall evaluate it as a claim.
New methodologies for historical linguistics?

Figure 1. Topological mismatches between 1A. Blust (2009) and 1B. Gray et al. (2009). The numbers on nodes refer to the subgroups discussed in Table 1 of the supplementary materials, following Blust (2009). The different subgroups are: 1: Polynesian, 2: Central Pacific, 3: Nuclear Micronesian, 4: South-East Solomonic, 5: North-Central Vanuatu, 6: South Vanuatu, 7: New Caledonia/Loyalty islands, 8: North New Guinea, 9: Papuan Tip, 10: Meso-Melanesian, 11: Admiralties, 12: St. Matthias, 13: Western Oceanic, 14: Oceanic, 15: South Halmahera/West New Guinea, 16: Eastern Malayo-Polynesian, 17: Central Malayo-Polynesian, 18: Central-Eastern Malayo-Polynesian, 19: Central Maluku, 20: Yamdena-North Bomberai, 21: Celebic, 22: Greater South Sulawesi, 23: Greater Barito, 24: Malayo-Chamic, 25: Barrier Islands/North Sumatra, 26: North Sarawak, 27: North Borneo, 28: Philippines, 29: Greater Central Philippines, 30: Western Malayo-Polynesian, 31: Malayo-Polynesian, 34: Malayo-Sumbawan. The convention [ ] indicates that part of a subgroup identified in Figure 1A can be found at that node; ( ) indicates that a single member of the identified subgroup is present at that node. Unnumbered nodes in the tree in 1A represent nodes discussed in the text of Blust (2009), but not included in Gray et al. (2009). Unnumbered nodes in 1B do not correspond to nodes in Blust (2009), and in the main represent subgroups that are not recognised for the Austronesian family.
data, such as the three spinal nodes between Oceanic and Central Pacific (Central Pacific is considered as representing a first-order subgroup of Oceanic). We illustrate the lack of replication in a detailed examination of Gray et al.’s technique with the final node in their tree, Polynesian, since it is the least controversial subgroup within Austronesian, and has attracted the most historical work.

3. **Polynesian**

The Polynesian subgroup is successfully replicated *in toto* by Gray et al. (2009, Supporting Online Materials), and with Fijian and Rotuman accurately grouped together as the Central Pacific subgroup of Oceanic. While the Polynesian subgroup is replicated to the exclusion of other languages, its internal structure bears no resemblance to the trees proposed by Marck (2000) or Pawley (1972). Although Polynesian is perhaps the most studied and best-understood part of the Austronesian tree, Gray et al.’s results are at odds with accepted trees for Polynesian, so that almost none of the internal subgroups are replicated. Figure 2 compares the structure of the subgroup in Gray et al. on the left, with a more accepted subgrouping (following Marck 2000) on the right, keeping the ordering of the languages identical. In this way the crossed lines on the right of the diagram indicate where Gray et al.’s methodology did not replicate the Polynesian family tree. The colours shown behind each language name reflect the subgroups accepted from Marck’s work, so that the Tongic subgroup, for instance, is shown in dark blue behind Tongan and Niuean. As can be seen in Table 2, while ‘rakes’ are not falsely replicated, all but one of the Polynesian subgroups are scrambled. The overall rate of successful replication is 83%; this might sound highly successful, but this cannot be assessed until we calibrate it by comparison with an attempted replication of a more familiar tree, which we present below.

Figure 3 shows a clustering of those languages of Europe that are most fully represented in the *World Atlas of Language Structures* (Haspelmath et al. 2005); the tree is drawn by reference to the typological features coded in WALS. We use this merely as an objective way to produce a computational clustering not based on the comparative method. Overall, Figure 3 replicates the six European subgroups of Indo-European at an 84% rate (Donohue et al. 2011), representing a closer replication of generally accepted subgrouping than Gray et al.’s (2009, Supporting Online Materials) replication of Polynesian. Nonetheless, it clearly does not represent a satisfactory replication of the accepted subgroup relationships of (Indo-)European languages drawn using the comparative method. Where subgroups are successfully reproduced they represent geographically contiguous blocks of languages, and wherever we find (from a subgrouping perspective) surprising clusters a
New methodologies for historical linguistics?

Figure 2. Polynesian phylogenies compared: 2A. A comparison of Gray et al.’s (2009) clustering, on the left, with the accepted subgrouping of Marck (2000) shown on the right. The figure uses Gray et al.’s 2009 clusters, but colour-codes subgroups according to Marck (2000) to highlight the lack of fit between the two trees. 2B. A map illustrating the geographically-based misclassification of Polynesian languages by Gray et al. (2009).
<table>
<thead>
<tr>
<th>Subgroup:</th>
<th>Contains:</th>
<th>Closest replication:</th>
<th>Problems:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tongic</td>
<td>Tongan, Niuean</td>
<td>Tongan-Samoan (93%)</td>
<td>Does not include Niuean; falsely includes Samoan</td>
</tr>
<tr>
<td>Nuclear Polynesian</td>
<td>(all but Tongic)</td>
<td>(all but Tongan and Samoan) (63%)</td>
<td>Does not include Samoan; falsely includes Niuean</td>
</tr>
<tr>
<td>Ellicean</td>
<td>Samoan, Ellicean Outliers and Eastern Polynesian</td>
<td>(67%)</td>
<td>Falsely includes Tikopia; does not include Samoan</td>
</tr>
<tr>
<td>Ellicean Outliers</td>
<td>Nukuoro, Kapingamarangi, Takuu, Sikaiana, Luangiua, Tokelau and Tuvalu</td>
<td>Tikopia, Nukuoro, Kapingamarangi, Takuu, Sikaiana and Luangiua (58%)</td>
<td>Falsely includes Tikopia; does not include Tuvalu and Tokelau.</td>
</tr>
<tr>
<td>Eastern Polynesian</td>
<td>Rapa Nui, Mangareva, Marquesan, Hawaiian, Tuamotuan, Rarotongan, Penrhyn, Maori, Tahitian and Rurutuan</td>
<td>(identical) (100%)</td>
<td>Perfect replication</td>
</tr>
<tr>
<td>Central-Eastern Polynesian</td>
<td>Mangareva, Marquesan, Hawaiian, Tuamotuan, Rarotongan, Penrhyn, Maori, Tahitian and Rurutuan</td>
<td>Hawaiian, Tuamotuan, Rarotongan, Penrhyn, Maori, Tahitian and Rurutuan (98%)</td>
<td>Does not include Mangareva and Marquesan.</td>
</tr>
<tr>
<td>or</td>
<td>Rapa Nui, Mangareva, Marquesan, Hawaiian, Tuamotuan, Rarotongan, Penrhyn, Maori, Tahitian and Rurutuan</td>
<td>Hawaiian, Tuamotuan, Rarotongan, Penrhyn, Maori, Tahitian and Rurutuan (100%)</td>
<td>Falsely includes Rapa Nui.</td>
</tr>
<tr>
<td>Marquesan</td>
<td>Mangareva, Marquesan and Hawaiian</td>
<td>Mangareva and Marquesan (81%)</td>
<td>Does not include Hawaiian</td>
</tr>
<tr>
<td>Tahitic</td>
<td>Tuamotu, Rarotongan, Penrhyn, Maori, Tahitian and Rurutuan</td>
<td>Tuamotu, Rarotongan, Penrhyn, Hawaiian, Maori, Tahitian and Rurutuan (75%)</td>
<td>Falsely includes Hawaiian</td>
</tr>
</tbody>
</table>
geographic explanation is readily available. Romanian is removed from the rest of Romance, and clustered with the geographically proximate Albanian; Bulgarian is separated from its Balto-Slavic relatives and placed with neighbouring Greek, with which it shares much of its recent history (no single pan-Balkan group emerges due to word-order differences and limited case marking for Greek and Bulgarian, the combined effect of which pulls them into a Celtic cluster); geographically and socially distant Icelandic is separate from the other Germanic languages, reflecting its relative isolation for the last millennium; and English is grouped with Norwegian and Swedish, not its West Germanic kin on mainland Europe, reflecting the extensive Scandinavian influence on English. The geographic explanations unsurprisingly indicate, following Holman et al. (2007), that typological features can diffuse across space.

Keeping the European calibration study in mind, it is not hard to account for discrepancies between the two trees in figure 2a.

- The inaccurate subgrouping of Tongan with Samoan reflects the long-acknowledged lexical conservatism of these two geographically proximal languages, which has resulted in a higher-than-normal percentage of lexical retentions from Proto-Polynesian (e.g., Biggs & Clark 2000, Greenhill et al. 2008), as well as the strong and extended contacts between these two island groups (e.g., Kirch & Green 2001).
- Niuean, on the other hand, has been extensively lexically influenced by Nuclear Polynesian languages, and is inaccurately subgrouped with them.
- Tikopia is the most westerly of the ‘Futunic outliers’, first order descendants of Proto-Nuclear Polynesian, and has most of its external contacts, including

Figure 3. European language subgroups replicated at an 84% rate.
lexical borrowings, with Ellicean Outlier languages, resulting in its incorrect placement.

– Tuvalu and Tokelauan, on the other hand, are much more easterly (see Figure 2B), and have been extensively lexically influenced through contact with Samoan and Eastern Polynesian languages.

– Within Eastern Polynesian the isolate Rapa Nui is incorrectly grouped with Marquesan and Mangareva, reflecting the only (relatively) close associations available to the geographically isolated Easter Islanders.

– Marquesan and Mangareva are thus falsely subgrouped away from their sister-language Hawaiian, which is combined into a subgroup that otherwise replicates Tahitic.

In all cases, the differences between the traditional tree and Gray et al.’s are due to one or a combination of: (a) lexical borrowing between geographically and socially proximal languages (e.g., Niuean and East Uvean); (b) the association of a socially and geographically isolated language with a nearest neighbor (e.g., Rapa Nui); or (c) a classification based on shared retentions (Samoan and Tongan). Figure 2B shows the locations of the different Polynesian languages; the numbers refer to the position from the comparative tree in Figure 2, counting from the top of the figure (except that ‘Futuna’ and ‘West Futuna’, which represent the same language, are ‘both’ coded with the same numeral ‘19’). Comparing the locations of the languages in Figure 2B with the degree to which the different Polynesian subgroups have been successfully replicated (Figure 2A and Table 2), all cases of successful reproduction are in geographically uncomplicated regions (especially the eastern Pacific), where there is less chance of ‘lexical contamination’ due to contact with non-Polynesian languages. Conversely, the subgroups which are least well replicated (e.g., Nuclear Polynesian, Ellicean, Ellicean Outliers) are all (wholly or partly) in socially complex regions, in the centre or the west of the Polynesian world, involving contact between Polynesian and non-Polynesian languages. Clearly geography, at least as much as known linguistic subgrouping, is a predictor for the results of Gray et al.’s clustering analysis based on lexical identity.

4. Analysis

Individual words, the quanta measured by Gray et al.’s technique, are the elements most susceptible to diffusion or replacement by diffusion (Séguy 1971, Hess 1979, Goeb 1984, Thomason & Kaufman 1988, Kessler 2001, others). Consequently, following McMahon & McMahon (2006), a lexical method, blind to the difference between inherited retentions, shared innovations and areal diffusions, will
be influenced by much “undetected borrowing between neighboring languages” (Gray et al. 2009:480). For this reason most linguists do not consider phylogenies based solely on lexical methods to have great value in reconstructing linguistic ancestry. For example, 18.5% of the basic vocabulary in modern English (from a 200-item Swadesh wordlist) has been borrowed (see supplementary materials), a result consistent with studies of other languages (Séguy 1971, Goebel 1984, Kessler 2001), and very close to the 20% level that Gray et al. cite as likely to be problematic for their method (2009:479). There is no reason to believe that these figures are unusually high.

Studies of basic vocabulary in some of the Austronesian languages considered by Gray et al. show up to 48% borrowing rates (e.g., Grant 2005). This emphasises that lexical cognate data, even from ‘basic’ wordlists and organised into cognate sets by experts, are not a reliable indication of linguistic ancestry. For that reason the comparative method, relying on sound-meaning pairings and regular correspondences, and clustering only on the basis of shared innovations, is considered the ‘gold standard’ in establishing relationships. Shared innovations, and not common retentions (including commonly retained words), are the sole criterion that can establish subgroupings which are valid representatives of shared ancestry. Individual words are subject to borrowing, and this effect is apparent in Gray et al.’s results: the languages of eastern Indonesia group with other languages in the east; the socially and geographically distant Chamic languages of mainland Southeast Asia are incorrectly classed as a first-order subgroup of the Malayo-Sumbawan group in the west; the Eastern Polynesian languages, lacking any non-Polynesian social neighbors, are correctly grouped together; and the isolate (within Malayo-Polynesian) Javanese is incorrectly embedded in a subgroup that contains the languages that surround it, geographically and socially. Other examples of misclassification occur whenever the ‘comparative’ family tree does not match the socio-geographic distribution of languages listed in our supplementary materials.

Where distance and subgrouping coincide, Gray et al. (2009) replicate subgroups identified in Blust (2009); where they do not, the phylogenies diverge (see Donohue, Wichmann & Albu 2008 on phylogenetic analysis based on abstract typological features). Isolates are assigned to subgroupings with geographically proximal languages when they are not closely related. For example, Javanese is surrounded by languages of the Malayo-Sumbawan group and has been incorrectly included in that subgroup by Gray et al., although it is clearly not part of this subgroup (Adelaar 2005b, Blust 2009). While we can explain this discrepancy as due to borrowing from the prestigious Javanese into its close neighbors, we are left with the conclusion that diffusion, rather than inheritance, is the fundamental signal detected by Gray et al.’s method. We have shown that lexical comparison of this sort, paying no attention to change between cognate forms, is not a proxy
for the comparative method.6 We have conclusively shown that whenever social distance and linguistic subgrouping would be expected to yield different clusters, the lexical comparison reflects clusters-by-distance more accurately than linguistic subgroups, a result presaged by discussion in Gray & Atkinson (2003). This means that the technique is a valuable tool for use in detecting contact, but not for replicating models of or forming hypotheses about linguistic origins.

While binary splits in trees are an unavoidable result of the algorithms used, the vast number of intermediate subgroups has strong implications for the validity of the technique that Gray et al. use to calibrate and date linguistic splits, one of the main stated aims of Gray et al. (2009). The technique of lexical comparison practised by Gray et al. is in some ways reminiscent of Greenberg’s ‘mass comparison’ (Greenberg 1987a, b), and most of the criticisms that have been leveled against that technique (e.g., Chafe 1987, Goddard 1987, Campbell 1988, Adelaar 1989, Matisoff 1990) apply to Gray et al.’s work as well. Gray et al. employ stricter controls for judging cognacy, but do not distinguish innovation from retention and inheritance from diffusion, given the lack of expert sociolinguistic and comparative knowledge in many (if not most) of the regions from which their wordlists are taken, and do not take account of chance similarities, onomatopoeia, and sound symbolism.

5. Conclusions

Gray et al.’s (2009) technique of cognate comparison does not adequately consider non-phylogenetic explanations, specifically the existence of shared retentions and the effects of lexical diffusion, as likely explanatory factors. That both linguistic phylogenies and lexical diffusion typically operate across contiguous regions means that only by examining the technique’s success at replicating non-contiguous subgroups can we find a reliable sounding board to evaluate this innovative methodology. In our applications of this heuristic, Gray et al.’s methodology preferentially detects geography/social contact. Similarly, while alternative methodologies generate valuable new ideas, claims of accuracy must rely on the replication of established groups and cannot claim that the generation of speculative new proposals is proof of matching results. Claims must also consider non-phylogenetic explanations, such as borrowing, even in the face of an overall close

---

6. This is not to dismiss the value of all lexical methods; an empirically-determined list of meanings that are maximally resistant to borrowing (Holman et al. 2008a, 2008b), taking into account a measure of ‘phonological distance’ between forms, allows surprisingly good results in replicating a number of language families.
match. This is especially important when the lexical data that generate these claims are susceptible to social and spatial proximity at least as much as inheritance.

References


Résumé

Des études récentes affirment que des observations fondées exclusivement sur le lexique peuvent émuler les résultats de la linguistique historique classique. Nous passons ainsi en revue une étude portant sur la famille austronésienne (Gray et al. 2009), et y trouvons deux problèmes principaux. D’une part, cette classification fondée sur le lexique omet de distinguer entre le partage d’innovations et le partage de traits hérités de la langue-mère. D’autre part, dans la mesure où les langues et familles linguistiques tendent à se disperser dans des zones contiguës, il en découle logiquement que la géographie — ou la distance sociale — correspond le plus souvent, pour l’essentiel, à la phylogénie. Gray et al. (2009) ne réussissent pas à reproduire tout l’arbre phylogénétique, à cause de phénomènes d’emprunts. Or, chaque fois que la phylogénie ne se superpose pas à une simple expansion géographique, il est possible de définir si la méthode de Gray et al. détecte la phylogénie ou la géographie. Dans tous les cas, lorsque phylogénie et géographie ne se superposent pas, on constate que les regroupements de langues proposés par Gray et al. reposent sur la géographie humaine (la distance sociale) et non l’histoire phylogénétique. En somme, l’observation des apparentements lexicaux ne saurait remplacer la méthode comparative classique.

Zusammenfassung

Authors’ Addresses

Mark Donohue
Linguistics/College of Asia and the Pacific
The Australian National University
Canberra, Australia ACT 0200
mark.donohue@anu.edu.au

Tim Denham
Archaeology Program/School of Historical and European Studies
La Trobe University
Melbourne, Australia 3086
t.denham@latrobe.edu.au

Stephen Oppenheimer
Institute of Cognitive and Evolutionary Anthropology/School of Anthropology
University of Oxford
64 Banbury Road
Oxford, United Kingdom OX2
stephen.oppenheimer@anthro.ox.ac.uk