BRILL

Language Dynamics and Change 2 (2012) 1–27

brill.com/ldc

Internal classification of the Alor-Pantar language family using computational methods applied to the lexicon

Laura C. Robinson[a], Gary Holton[b]
a) Postdoctoral researcher, University of Alaska Fairbanks
Corresponding author
lcrobinson@alaska.edu

b) Professor of Linguistics,
University of Alaska Fairbanks
~~*Director, Alaska Native Language Archive*~~
gmholton@alaska.edu

Abstract
The non-Austronesian languages of Alor and Pantar in eastern Indonesia have been shown to be genetically related using the comparative method, but the identified phonological innovations are typologically common and do not delineate neat subgroups. We apply computational methods to recently collected lexical data and are able to identify subgroups based on the lexicon. Crucially, the lexical data are coded for cognacy based on identified phonological innovations. This methodology can succeed even where phonological innovations themselves fail to identify subgroups, showing that computational methods using lexical data can be a powerful tool supplementing the comparative method.

Keywords
Alor; Panta; subgrouping; split decomposition; phylogenetics; language classification

1. Introduction

The standard method for establishing internal genealogical relationships within a language family relies on tracing shared innovations, usually phonological ones. In many language families, however, continued contact and convergence renders the family tree model inadequate for the purposes of classification (cf. Krauss, 1973 on Athabaskan; Sidwell, 2009 on Austroasiatic). In this paper we discuss one such problematic case from the Alor-Pantar (AP) family of eastern Indonesia and propose an alternate method of classification, employing computational phylogenetic methods applied to lexical data.

 DOI:

Our approach differs from previous applications of computational methods to linguistic data in several ways. First, unlike Dunn et al. (2008), we use lexical rather than typological features. The use of typological features, even large numbers of typological features, has been criticized because they are considered extremely susceptible to borrowing (see Donohue et al., 2011). Second, unlike lexicostatistical models which rely on subjective lexical similarity judgments, we use the comparative method to identify cognate sets, as in Gray et al. (2010). All lexical items counted as cognates in our database are rigorously vetted using the known sound correspondences from Holton et al. (2012), hereafter H2012. In this way, we have attempted to mitigate the subjectivity that was pervasive in lexicostatistics. Also in contrast to traditional lexicostatistical methods, our computational models account for patterns of shared cognacy among languages as opposed to just counting raw cognacy percentages.

Third, rather than examining large families such as Austronesian (Gray and Jordan, 2000) or Indo-European (Gray and Atkinson, 2003) we restrict our focus to one very small and well-defined family. The Alor-Pantar family has approximately twenty-one languages, and we include data from twelve of those languages in our database. We chose this subset intentionally in order to match the set of languages used in the reconstruction of proto-Alor-Pantar (pAP) in H2012. The genealogical unity of the Alor-Pantar family was demonstrated by Stokhof (1975) using lexicostatistics, and it has recently been confirmed by means of the comparative method (H2012). Using data from a relatively small and well-defined language family has the advantage of restricting the uncertainty in the results to the question at hand: namely, the internal classification of the languages. We do not need to determine if the languages we have chosen indeed form a genealogical unit or if an unrelated language has accidentally been included.

Finally, rather than examining potential deeper affiliations, we restrict our attention to a single family. While the wider genealogical affiliations of the Alor-Pantar family are frequently alluded to, much of this work has been based on a dearth of linguistic data and is necessarily speculative (though see Robinson and Holton, to appear). By restricting our attention to a single, well-defined family, we are able to draw on a controlled, high-quality lexical data set.

2. Previous Attempts at Subgrouping

The Alor-Pantar family itself is unique in being one of only two pockets of non-Austronesian languages in Island Southeast Asia west of mainland New

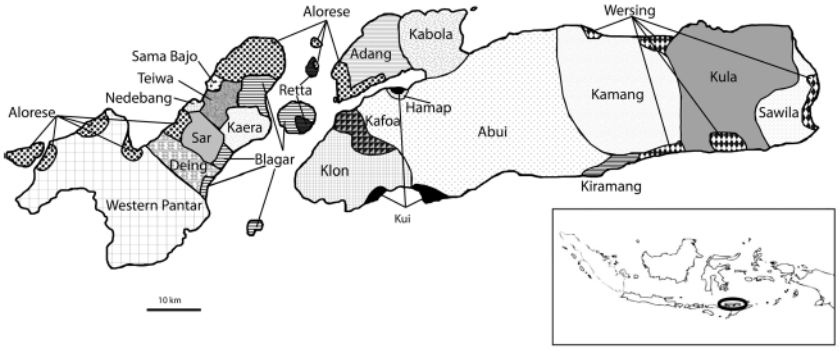

Figure 1. Map of the Alor-Pantar languages

Guinea.[1] The distinctive typological profile of the languages of the Alor-Pantar archipelago (defined as the islands of Alor and Pantar together with the small islands in the intervening Pantar Strait) and that of several languages on neighboring Timor and Kisar was first recognized in the literature roughly a century ago. Early Dutch military reports focused on racial and cultural distinctions (Anonymous, 1914), and by the mid-twentieth century at least one language of Alor was recognized as being non-Austronesian in character (Nicolspeyer, 1940). Survey work conducted in the early 1970s made clear the existence of perhaps two dozen obviously related non-Austronesian languages in this region (Fig. 1), but internal classification remained speculative, owing to a lack of primary data (Stokhof, 1975).

The recent availability of new lexical data for these languages facilitates a more robust attempt at subgrouping, making it possible for the first time to examine both the internal and external linguistic relations of the Alor-Pantar languages. Stokhof attempted to subgroup these languages using lexicostatistical methods, but only with the availability of new data in the twenty-first century was it possible to apply the standard techniques of the comparative method and attempt a reconstruction of the proto-Alor-Pantar language (H2012).

The task of language classification involves determining which languages share a common ancestor. Two languages can be classified as belonging to the same family when they can be shown to descend from a common ancestor through shared basic vocabulary, shared grammatical morphology, and a set of regular sound changes (Campbell and Poser, 2008). The task of subgrouping is a finer-

[1] Donohue (2007) argues that the now extinct language Tambora, located some 700 km west of Alor-Pantar, may also have been non-Austronesian.

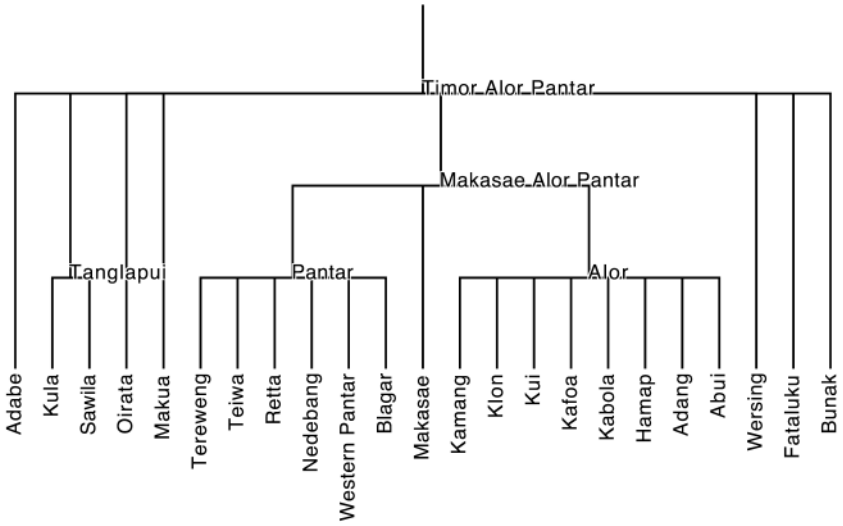

Figure 2. Timor-Alor-Pantar classification from Ethnologue 2005 (Gordon, 2005)

grained one which attempts to determine which languages within a language family are more closely related than others. Borrowing an analogy from biology, lions, tigers, and bears all belong to the same family of mammals, but lions and tigers subgroup together more closely as part of the feline group. Within linguistics, subgrouping typically relies on the same methodology as that employed to show two languages are related. If two languages can be shown to be related—that is, if they can be shown to descend from a common ancestor through shared basic vocabulary, shared grammatical morphology, and regular (sound) changes—then they can be said to subgroup together if they share a subset of those changes. While innovations can occur independently in two languages, the greater the number of changes considered and the more unique and unusual those changes, the less likely they are to have occurred independently in the two languages, and thus the more likely they are to provide evidence for a shared history.

This method of subgrouping is complicated by the fact that it requires *a priori* application of the comparative method to identify the relevant sound changes and lexical or grammatical innovations which have occurred in the history of a language family. In the case of the Alor-Pantar languages, this has only recently been accomplished (H2012). In the absence of the comparative method, linguists have often resorted to more impressionistic and subjective methods for subgrouping languages. The Alor-Pantar languages have not been immune to such ad-hoc methods. In Figs 2 and 3 ~~below,~~ we reproduce family tree diagrams

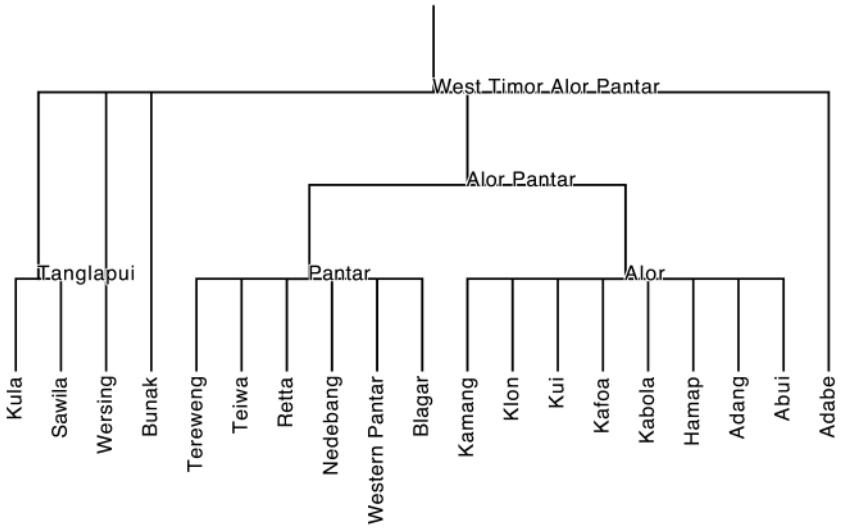

Figure 3. Timor-Alor-Pantar classification from Ethnologue 2009 (Lewis, 2009)

for the Timor-Alor-Pantar language family from two recent editions of the *Ethnologue*, a widely referenced catalog of the world's language families (Gordon, 2005; Lewis, 2009). The intermediate nodes on the trees represent subgroups of languages which are asserted to be more closely related to one another. Oirata, Makua, Makasae, Fataluku, and Bunak are all languages of Timor. Adabe is spoken on an island between Alor and Timor, but it is actually Austronesian and therefore does not belong in either tree.[2]

Though the two versions of the family tree differ in certain respects, overall they are quite similar. Looking more closely at the trees, we can discern two primary factors driving the subgrouping process. The first factor is geographic. The nodes labeled Pantar and Alor consist of languages spoken on Pantar and Alor islands, respectively. Membership in these two subgroups remains stable in the two trees. The second factor driving the subgrouping process in the *Ethnologue* is typological. Kula and Sawila (which as late as 1996 were viewed as a single language) are classified as a distinct subgroup, in spite of the fact that they are spoken on Alor island and hence group geographically with the other Alor languages.

---

[2] The most recent published survey of languages of East Timor lists three mutually intelligible dialects of Adabe, namely: Rahesuk, Resuk, and Raklungu (Hajek, 2010). No direct information regarding genetic affiliation is given, though these Adabe dialects are not included among the listing of four non-Austronesian languages of East Timor; so presumably these dialects are also Austronesian.

Table 1. Sound changes found in at least two languages (H2012).[3]

Change	Languages
*b > f	Tw, Nd, Ab (in Tw and Nd only non-initially)
*b > p	Km, Sw, We
*d > r	Ab, Ki (in Ki only finally)
*g > ʔ	Bl, Ad
*k > Ø / _#	Bl, Ad
*q > k	WP, Bl, Ad, Kl, Ki, Ab, Km, Sw, We (Ad ʔ < k < *q)
*s > h	Bl, Ad, Kl
*s > t	Ab, Sw, We
*h > Ø	everywhere but Tw and WP
*m > ŋ / _#	WP, Bl, Ad
*n > ŋ / _#	Nd, Ke, WP, Bl, Ad, Ab, Km, Sw, We
*l > i / _#	Tw, Ke, Ad, Km
*l > Ø / _#	Nd, WP, Ab
*r > l / V_V	Nd, WP, Ad, Km
*r > Ø / _#	Tw, Ke, WP
*r > i / _#	Bl, Ki, Ab

This is likely due to the fact that these languages are lexically and morphologically somewhat different in character than the surrounding languages. For example, grammatical relations in Kula have been described as behaving according to an inverse system not found in the other languages of Alor (Donohue, 1996). Similarly, typologically distinct Wersing is placed in a subgroup of its own as a family-level isolate.

A third factor driving subgrouping at this level can be labeled impressionistic. The changes between the 2005 and 2009 versions of the *Ethnologue* tree reflect, among other things, reassessment of the position of the Timor~~–Kisar~~ languages. In the 2005 version, the Timor language Makasae is coordinate with the Pantar and Alor subgroups, while the remaining Timor languages Oirata, Makua, Fataluku, Bunak, and Adabe are listed as family-level isolates. In the 2009 version only Bunak and Adabe remain. Given that reconstruction of proto-Timor-Alor-Pantar has not yet been completed (though see Schapper et al., 2012), this reclassification of the subgroups must necessarily be based on impressionistic evidence. It may well be based at least in part on geography, as indicated by the renaming of Timor-Alor-Pantar to West Timor-Alor-Pantar, since the 2009 classification includes only the westernmost of the Timor languages. This type of subjective subgrouping is not at all unusual and occurs widely throughout the field. Combined with the geographic and typological factors noted above, the impressionistic method remains a frequently employed technique for determining internal linguistic relationships within established language families.

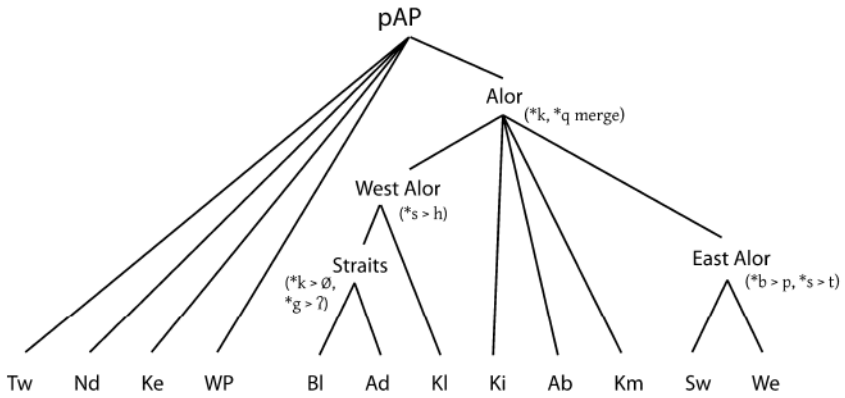

Figure 4. Subgrouping of Alor-Pantar based on shared phonological innovations (H2012)

With this background it is natural to ask why linguists choose to rely on such ad-hoc methodology for determining linguistic subgroups. Of course the primary answer is that, in many cases, the work of the comparative method remains to be completed, so we do not have knowledge of regular historical changes which could be used to discern shared linguistic history. But even when we do have knowledge of historical sound changes, it is not always easy to determine subgroups with a high degree of certainty. Rather than proceeding in a neat hierarchical fashion, sound changes often cross-cut each other in a wave-like pattern that is not well-described by a tree. This is the case in the history of the Alor-Pantar languages. While it is possible to identify many phonological innovations in the history of Alor-Pantar, only two of those innovations occur in the same subset of languages (*g > ʔ and *k > Ø / _#, both in Blagar and Adang). Each of the other phonological innovations delineates a distinct subset of languages. Table 1 lists all of the sound changes which have occurred in at least two of the daughter languages. Changes which are restricted to final (_#) or medial (V_V) position are followed by a slash and an indication of which position they occur in.

H2012 use a subset of these changes to delineate four subgroups within the Alor-Pantar family (see Fig. 4). An Alor subgroup is defined by the merger of *k and *q. Within the Alor subgroup, a West Alor subgroup is defined by the change *s > h, and an East Alor subgroup is defined by the two changes *b > p and *s > t. The former change is also shared with Kamang (Km); the latter with Abui (Ab). So while it is tempting to expand this group, only Sawila (Sw) and Wersing (We) share both of these innovations, defining the subgroup we refer to as East Alor. Finally, within West Alor, a Straits subgroup is defined by the

Table 2. Forms for 'sharp' and expected reflexes of pAP medial *g.

	'sharp'	medial *g
Teiwa	mahan	h
Nedebang	maxaŋ	x
Kaera	magaŋ	g
W Pantar	maggaŋ	gg
Blagar	maŋ	Ø
Klon	maŋ	g
Kui	maŋan	g

changes *k > Ø, *g > ʔ, and *s > h. Only the latter change is shared with Klon (Kl), providing weak support for the intermediate level of the grouping labeled West Alor. The remaining changes cross-cut these and do not provide additional subgrouping information. Mapping these changes results in the ~~following~~ family tree diagram.

While this procedure does create a satisfactory tree structure, the method is not very robust because it relies heavily on the subjective choices of which sound changes are used to delineate subgroups. Further, the tree as drawn obscures much of the shared history within the family. For example, one of the changes defining the East Alor group is also shared with Kamang (Km), while the other is also shared with Abui (Ab), though neither Kamang nor Abui reflects both of these changes. ~~This~~ fact that many of the changes cross-cut each other is a general problem with the tree-drawing method. For example, the change *s > h is used to group Adang with the other West Alor languages in Fig. 4. But Adang could equally be grouped with Western Pantar (WP) and Blagar (Bl) based on the shared change *m > ŋ (see Table 1 above). The task of subgrouping asks us to produce a single family tree, when in fact the evidence argues for many different and possibly conflicting trees.

3. The Lexical Data

While the phonological innovations identified for Alor-Pantar by the comparative method do not yield neat subgroups, they do allow us to distinguish inherited from borrowed forms (as long as the borrowings are relatively recent). We consider forms to be non-cognate (and thus potentially borrowed) when the sound correspondences are not the ones predicted by the comparative method. Cursory examination of the forms for 'sharp' in Table 2 reveals the appearance of potentially cognate forms.

The words are indeed all very similar, beginning with a labial nasal and ending with another nasal. However, neither the Klon nor the Kui form for 'sharp' has

Table 3. Forms for 'small' and expected reflexes of pAP *d.

	'small'	*d
Kui	kadin	d
Abui	kidiŋ	r
Kamang	kidiŋ	t

the velar stop reflex of original medial *g that would be expected based on the comparative method, and extra segments of unknown origin have been added to the Kui form. In other words, this word does not obey the rules of regular sound change in Klon and Kui. Thus, it cannot be an inherited lexical item in Klon and Kui but rather may be a borrowing from Blagar, a language which loses original medial *g entirely. The extra segments in Kui may have been added for independent reasons not explored here.

The data in Table 3 illustrate a similar point. Since we expect *d > r in Abui and *d > t in Kamang, the Abui and Kamang forms for 'small' are anomalous. They cannot reflect shared descent. Rather, they may have been borrowed at some stage from Kui.

Applying this methodology across the lexicon allows us to distinguish true cognates from apparent cognates or "lookalikes." We then use these cognate classes combined with our knowledge of which items can be reconstructed to proto-Alor-Pantar to identify shared lexical innovations that can be used to identify subgroups. In this way, the use of lexical innovations to determine subgrouping differs from lexicostatistical methods (Swadesh, 1950): the latter rely on subjective assessments of lexical "similarity," rather than shared innovations which are identified as cognate on the basis of established regular sound correspondences.

The problem then becomes one of scale. Rather than using the 16 sound changes listed in Table 1 to create subgroups based on phonological innovations, we have 400 lexical items for twelve languages, each grouping into between one and eight cognate classes. Just like the phonological innovations, the lexical innovations are not neatly bundled but cross-cut each other, making it nearly impossible to determine subgroups by hand. Rather than attempting to recognize tree-like signals by hand, we apply computational phylogenetic techniques to tease out subgroups. These methods are particularly appropriate to modeling lexical history because they allow for the existence of both vertical (shared descent) and lateral (borrowing, innovation) transfer events. In terms of linguistic history, vertical transfer refers to inheritance of a word form through regular phonological changes, as in English *foot* < Old English *fōt. Lateral transfer typically refers to borrowing, as in the borrowing of English *devour* < Old French *devorer 'swal-

Table 4. Lexical data for 'short' and 'small' in four AP languages.

	'short'	'small'
Teiwa	tuk	sam
Kui	tuk	kadin
Abui	bui	kidiŋ
Kamang	manuk	kidiŋ

Table 5. Cognate classes for 'short' and 'small.'

	'short'	'small'
Teiwa	0	0
Kui	0	1
Abui	1	2
Kamang	2	3

low.' Both types of processes occur in the lexical history of languages, and phylogenetic techniques can model both types of events simultaneously.

Whereas traditional approaches to language classification assume a sequence of innovations (mutation events), the phylogenetic models employed here view language change as variation in character states. In the case of lexical data, the characters are lexical items and the character states are the cognate classes. Where all modern etyma of a particular lexical item are cognate, all languages (taxa) exhibit the same character value. Where new lexical items are innovated or borrowed, new character states arise which differ from the original state. To take a simple example, consider a subset of lexical data for the forms 'short' and 'small' for the four taxa Teiwa, Kui, Abui, and Kamang in Table 4.

We can assign these lexical forms to numbered cognate classes in a linguistically informed way, recognizing forms which are not reflexes of pAP based on the absence of regular sound correspondences, as discussed above. Rather than discarding these lexical items as not reflecting descent from the protolanguage, we retain them and assign them to a distinct cognate class. If we number cognate classes 0, 1, 2, etc., the lexical data in Table 4 can be rendered numerically as in Table 5 ~~below.~~ Recall that the Kui, Abui, and Kamang words for 'small' were not considered cognate because they did not have the appropriate sound correspondences, and so they receive different numbers in Table 5, indicating that they belong to different cognate classes.[4]

[4] While the form for 'small' is identical in Abui and Kamang, we cannot rule out the pos-

In this example, the character 'short' has three character states (which we label 0, 1, 2) and the character 'small' has four character states (0, 1, 2, 3). For a given lexical item, the number of character states will range between one and the number of languages (four in the example above; twelve in our entire dataset). Where an item is cognate in all languages in the sample, there will be only one character state.

As a source of lexical data, we started with the same 400-item word list used by H2012. Using this dataset has several advantages. First, although there are approximately twenty-one Alor-Pantar languages, the subset of twelve languages selected by H2012 is ideal because it represents a wide geographic sample of the Alor-Pantar languages. Second, thanks to the previous comparative work, the phonological innovations for each language in this set have already been identified, facilitating straightforward identification of cognate classes. Third, this 400-item list was specifically tailored to include vocabulary relevant to the Alor-Pantar region.

One disadvantage of starting with this list is that it was developed at least in part to avoid innovations. It is essentially an expansion of a Swadesh 200-item basic vocabulary list, augmented with items which are informative to the task of linguistic reconstruction of Alor-Pantar. In particular, some items are included simply because they reflect a certain proto-sound or because they are thought to have widely distributed reflexes in the daughter languages. Hence, this lexical dataset is far from a random sample of vocabulary and might well undercount incidence of borrowing and innovation within the lexicon.

From the 400-item list we removed obvious recent introductions (such as 'corn') and known loans from non-Alor-Pantar languages (such as proto-Austronesian *takaw 'to steal'). We also removed several items for which data were missing for more than half of the twelve languages in the sample or which were largely redundant (e.g., we only included 'dolphin' and not 'whale' because the two were the same for most languages). The remaining 351 lexical items were coded numerically for cognacy as described above. Crucially, detectable intra-family borrowings were coded as distinct cognate classes as described above. In addition to these twelve languages, we also included proto-Alor-Pantar as a distinct taxon, coding each of the 97 lexical items in the dataset for which pAP forms have been reconstructed. Each lexical item that is a regular reflex of a pAP reconstruction was coded as belonging to the same cognate class as the pAP reconstruction. This process resulted in a 13×351 matrix ($13 \times 351 = 4{,}563$ character states).

---

sibility that these represent independent parallel innovation. Hence, these items are coded as distinct cognate classes.

Table 6. Binary cognate classes for 'short.'

	short1	short2	short3
Teiwa	1	0	0
Kui	1	0	0
Abui	0	1	0
Kamang	0	0	1

In order to avoid scaling factors, this matrix was converted to binary coding, generating a distinct binary character for each combination of lexical item and cognate class. This ensures that all distances between cognate classes are treated equally. If we return to the 'short' example discussed above, the binary matrix for Table 5 is shown in Table 6.

The word 'short' had three cognate classes in this sample set (Table 5), and so we now have separate characters for each cognate class. The first column asks the question, "Does this language have a word cognate with Teiwa *tuk* 'short'?" Teiwa and Kui do and thus receive a 1, while Abui and Kamang do not and thus receive a 0. Note that although Abui and Kamang are both coded the same in this column, this does not imply that their 'short' words are cognate. Rather, their non-cognacy is encoded in the differing values for Abui and Kamang in the short2 and short3 columns. The second column asks the question, "Does this language have a word cognate with Abui *bui* 'short'?", and so on. Converting our results to binary characters in this way yielded a 13×2542 matrix of lexical character values ($13 \times 2542 = 33{,}046$ character states). This matrix served as the primary dataset for our analyses. In the following two sections we discuss the application of two different computational techniques: a network model intended to represent the non-treelike nature of the data, and a Bayesian tree model intended to pick out the most probable of all the possible trees.

4. Split Decomposition Network

The method of split decomposition partitions the languages into groups according to whether they share a particular character state or not (Bandelt and Dress, 1992). In our dataset this corresponds to whether a lexical item belongs to a given cognate class or not. If all these splits are compatible with each other, then the method generates a tree structure with each split corresponding to a branch in the tree (and some branches supported by multiple splits). Yet, as discussed above, the splits in our dataset are not all compatible. That is, the cognate sets delineated by some lexical items overlap with those delineated by other lexical items. Thus, it is not possible to build a tree from the splits based on all the lexical characters.

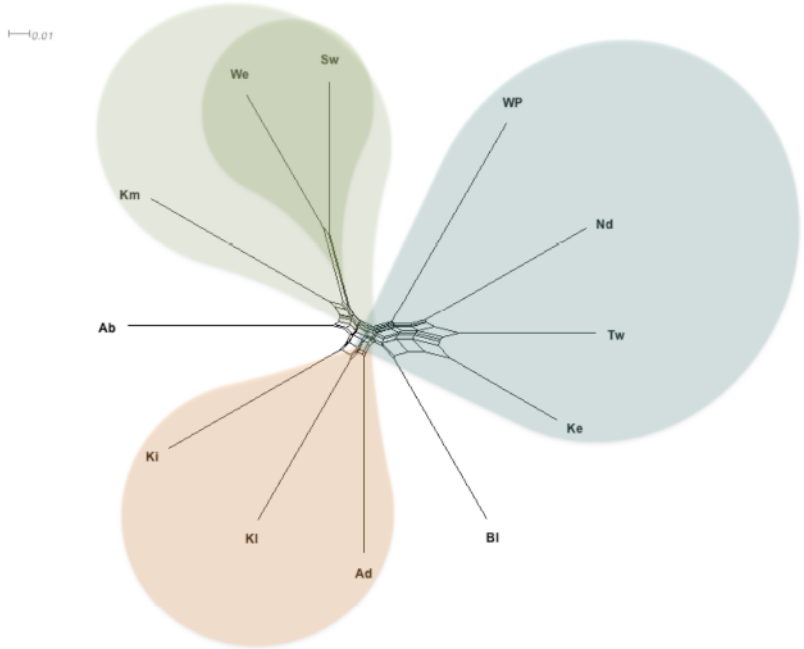

Figure 5. Split graph of NeighborNet analysis of lexical characters (excluding pAP)

Instead, we generate a network or split graph using the NeighborNet algorithm laid out in Huson and Bryant (2006). In a split graph, conflicting signals from the lexical data are represented as multiple edges (branches) connecting the language taxa. The greater the degree of multiple edges, or reticulations, present in the graph, the less treelike signal is present in the data. In particular, a split graph with no reticulations corresponds to a dataset with no conflicting character states and thus an entirely treelike signal.

We follow Gray et al. (2010) in using gene content distances as the distance metric in the NeighborNet analysis, since this metric most closely captures the unidirectional nature of lexical innovation (for details see Huson and Steel, 2004).[5] Reflexes within a particular cognate class arise through a single mutation event in the history of the language family, whereas lexical innovation may arise independently as distinct events in each of the daughter languages. The resulting split graph generated using the SplitsTree program (Huson and Bryant, 2006) is shown in Fig. 5. Note that, unlike family tree diagrams familiar in historical

5) However, we note that results using the uncorrected-P metric are largely similar.

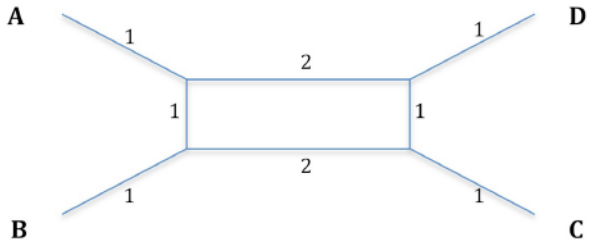

Figure 6. Idealized reticulate graph with four taxa A, B, C, D

linguistics, taxa (i.e., languages) in a split graph are not depicted as descending from a single ancestor node, but radiate out from a central point. This is because, although character states (i.e., cognate classes) are encoded in the database, the algorithm does not assume an ancestral state for each character and so cannot determine where to root the network. In this case, we have included pAP as another taxon with equal status to the other taxa, and as expected, the pAP node occurs in a central position within the graph, indicating that all the languages descend from a pAP node in the center of the graph. Excluding the pAP taxon does not significantly alter the shape of the network. In Fig. 5, the pAP node has been omitted for ease of representation.

Three primary regions can be identified in the graph, each separated by significant reticulation (webbing) at the center of the graph. An East Alor region groups Kamang, Wersing, and Sawila; a Central Alor region groups Kui, Klon, and Adang; and a Pantar region groups Kaera, Nedebang, Teiwa, and to a lesser extent Western Pantar. The high degree of reticulation within this latter group indicates a strong conflicting signal within this region. That is, of these three regions, the Pantar group is particularly non-treelike, suggesting a pattern of wavelike innovations in this region.

We can get a better handle on the degree of treelike signal present in the various regions by using the delta score (δ) generated by the SplitsTree program (version 4.12.6). The delta score indicates the degree of treelike signal present by measuring the extent to which distances between two taxa are additive (Holland et al., 2002), and it has been found in empirical studies to be an appropriate measure of reticulation in linguistic phylogenies (Wichmann et al., 2011). To see how delta is measured, consider the simple example of a graph with four taxa and edge (branch) lengths as specified in Fig. 6.

The distances AB and CD are each 3; the distance AD and BC are each 4; and the distances AC and BD are each 5. Summing on the latter two distances gives a greater value than summing on any other two pairs of distances within the graph. The delta score calculates the difference between the longest two pairwise sums,

Table 7. Delta scores for individual languages.

Pantar	Teiwa	0.25
	Kaera	0.28
	Nedebang	0.29
	Western Pantar	0.31
	Blagar	0.32
Central Alor	Adang	0.32
	Klon	0.30
	Kui	0.31
	Abui	0.33
East Alor	Kamang	0.26
	Wersing	0.24
	Sawila	0.24

normalized by the difference between the longest and shortest pairwise sums, thus yielding a number between 0 and 1.

$$\delta = \frac{(AC + BD) - (AD + BC)}{(AC + BD) - (AB + CD)}$$

For trees with no reticulation, the delta score will be zero since the longest two paths between pairs of nodes will be equal. Larger delta scores indicate greater divergence from treelike structure. For our hypothetical four-taxa example above, the delta score is 0.5, indicating a strong conflicting non-treelike signal. To calculate the delta score for a split graph with more than four taxa, we simply take the mean of the delta scores for each quartet (set of four taxa) containing the given taxon.

Applying this methodology to our dataset yields an average $\delta = 0.29$. This value is moderately high, reflecting the fact that, while some groupings do emerge in Fig. 5, there is significant reticulation between those groups. This figure can be contextualized by comparing with delta scores calculated by Gray et al. (2010) for Indo-European and Polynesian. Gray et al. used the same distance metric applied to basic vocabulary coded for cognacy, with borrowings retained; hence their data is largely comparable to our dataset for Alor-Pantar. What we find is that the delta score for the Alor-Pantar data lies midway between the more treelike score of 0.22 for Indo-European and the decidedly non-treelike score of 0.41 for Polynesian.[6] However, the mean for the dataset obscures considerable variation in the delta scores for the individual languages. As can be seen in Table 7,

[6] Wichmann et al. (2011: 216) derive a markedly higher value of $\delta = 0.39$ for the larger

the delta scores for the Alor-Pantar languages vary from a more Indo-European value of 0.24 to a more Polynesian value of 0.33.

The most treelike values are found in the East Alor grouping of Kamang, Wersing, and Sawila. The Pantar group of Teiwa, Kaera, and Nedebang has delta scores similar to the mean for the entire dataset; however, the value for Western Pantar is significantly higher, suggesting that similarities between Western Pantar and the remainder of the Pantar languages may be due more to borrowing than to shared descent. The delta scores for Adang, Klon, and Kui are typical for the dataset. An unexpected result of the graph in Fig. 5 is the position of Blagar as a relative isolate within the family. In contrast to the subgrouping based on the comparative method, Blagar groups not with Adang and Klon but rather with the Pantar languages—and then only weakly so. The relatively high delta score indicates a strong conflicting signal in the data. This will be discussed further in Section 6 below.

4.1. *Variation across the Lexicon*

The approach discussed above treats the lexicon as a monolithic whole. While we can recognize differences in the degree of treelike structure for each of the languages, we have no information about which parts of the lexicon are contributing to these differences. Yet there is no reason to expect all lexical items to have similar histories. For example, words for material culture may be borrowed more readily, as may words for introduced flora and fauna. More broadly, the patterns of lateral transfer may differ for different subsets of the lexicon. Here we make a preliminary attempt to tease out the contributions of various parts of the lexicon.

We begin by segmenting the lexical data into semantic categories as established in H2012. This semantic classification includes eleven categories, six of which are nominal and three of which are verbal. The distribution of lexical items from our database in each category is shown in Fig. 7 below. A plurality of items belongs to the verbal action/event category.

For each of these semantic categories, we generated a split graph using SplitsTree, restricting the data set to only those items belonging to the particular category. We then calculated delta scores for each of the 13 sample languages. This yielded a 11×13 matrix comparing the dimensions of semantic category and language. Since the delta scores correlate inversely with the degree of treelike structure present in the network for a given language, the matrix can be used to represent the variation of degree of treelike structure across the two variables of seman-

---

West Timor-Alor-Pantar. However, their results are not directly comparable to ours since they include the more distantly related Timor languages and are based on a smaller, 40-item word list.

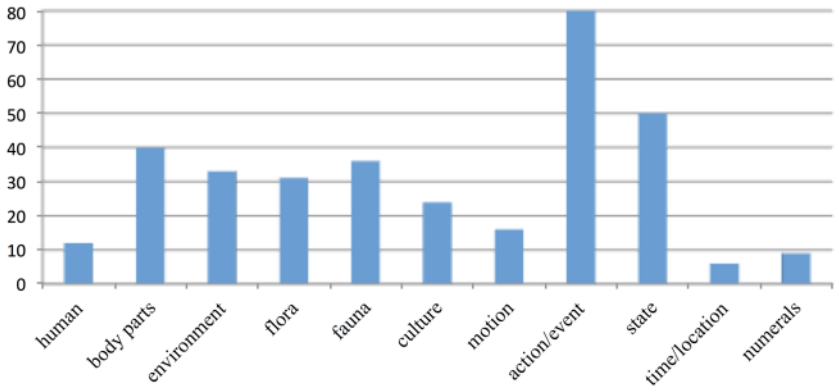

Figure 7. Distribution of lexical items by semantic category

tic category and language. Unfortunately, these variables are not sufficiently independent to admit a factor analysis ($p = 0.45$). However, we can extract principal components (PC) from the entire matrix, as shown in the biplot in Fig. 8. Principal component analysis tells us which factors are contributing to the overall structure of the results. The first principal component accounts for 60% of the variance, but all semantic categories contribute almost equally to this component. Only in the second principal component do we see some differentiation, with the time/location, human, and motion categories contributing slightly less to the treelike structure. However, the effect is small, since the second component contributes only 14% of the variance. Additionally, the divergence of PC2 for the time/location category may be explained in part by the relatively large proportion of absent lexical data within this category.

This preliminary analysis suggests that words denoting humans, motion, and time/location may be more stable across the Alor-Pantar languages. However, the effect is minimal and may be due to other factors not considered here. Crucially, our approach using ad-hoc semantic categories may miss wider patterns which cross the arbitrary semantic boundaries instantiated here. Ideally, we should be able to apply a factor analysis to the entire lexical dataset in order to determine the contribution of individual lexical items to the treelike signal in the data.

4.2. *Comparison to Phonological Innovations*

The splits graph based on lexical characters explicitly excludes known borrowings, since we coded intra-family borrowings as non-cognate whenever the correspondences do not adhere to previously established sound correspondences. This

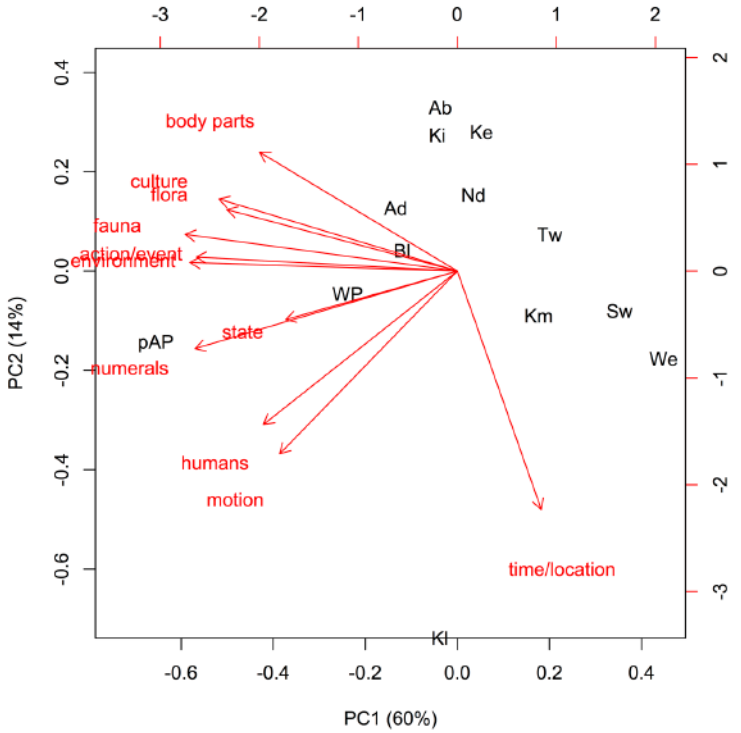

Figure 8. Biplot of Principal Components 1 and 2

still allows for the possibility that undetected borrowings may be present, leading to the observed reticulation in the splits graph. However, at least some of the reticulation must be due to wavelike innovations. This can be seen by comparing the splits graph based on lexical characters to one based on phonological innovations. While lateral transfer of lexical items does yield reticulate graphs, parallel phonological innovations and borrowing of sound changes can also generate incompatible splits. In presenting the subgrouping based on the comparative method, we noted a number of incompatible phonological innovations. The tree in Fig. 4 represents a compromise which privileges certain phonological innovations above others. For this reason, the traditional subgrouping methodology employed by H2012 is not directly comparable with an approach based on split decomposition of lexical characters. To facilitate a comparison, we apply split decomposition to the phonological characters determined by the comparative method.

Drawing on the reconstruction in H2012, we coded the reflexes of the fifteen proto-Alor-Pantar reconstructed consonants in initial, medial, and final position

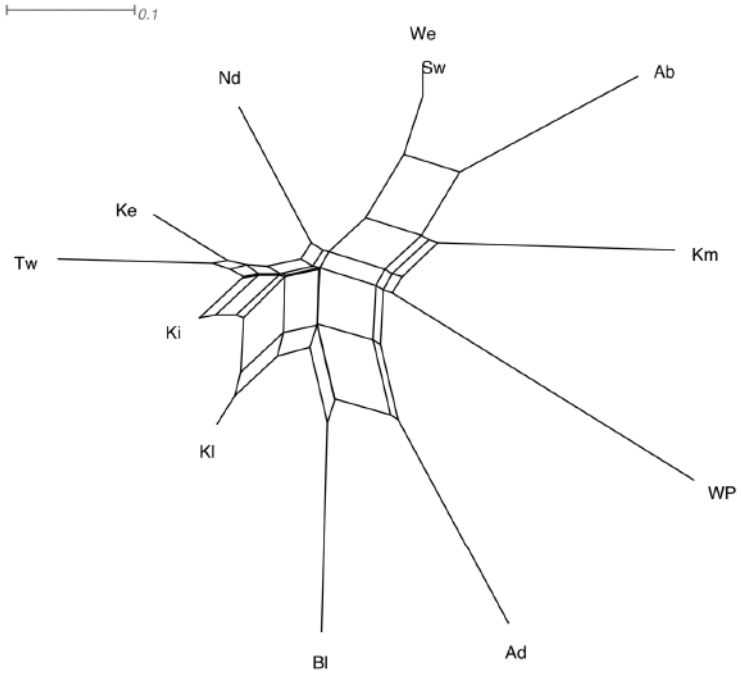

Figure 9. Split graph based on phonological innovations

in each of the twelve daughter languages considered in the lexical dataset, yielding 43 distinct characters. By focusing on reconstructed consonants we were able to ignore the effects of individual lexical items. In particular, neither borrowing nor irregular reflexes were accounted for in this dataset. The result of a NeighborNet analysis applied to these phonological data is given in Fig. 9.

Although the resulting split graph is based on phonological innovation determined using the comparative method, it shows even less treelike structure than the split graph based on lexical innovations (Fig. 5). This matches our intuition about the nature of the Alor-Pantar phonological innovations not delineating neat subgroups, as discussed in Section 2. The mean delta score for this phonological dataset is an extremely high 0.46, with a minimum value of 0.39 (for Kaera). The conflicting signal within the phonological data, therefore, is much stronger than for the lexical data, and fewer natural groupings emerge. The only grouping clearly shared between the graphs based on lexical and phonological characters is the tight grouping of the East Alor languages Wersing and Sawila, which, in fact, are co-terminus in the graph of the phonological innovations because they share all their phonological innovations (see Table 1). Other languages are sepa-

rated by significant amounts of reticulate structure, as might be expected with a dialect chain. This suggests that the sound changes may have diffused across the family—a not altogether unexpected result, given that many of the phonological innovations are extremely common cross-linguistically. We return to this point below, but first we consider an alternate methodology for detecting treelike signals in the lexical data.

5. Bayesian Tree Model

Split decomposition provides a convenient way of visualizing data which are not treelike, because it removes the requirement that the splits determined by individual characters are compatible with each other. An alternate approach instead attempts to find the tree with the best possible fit to the data. Where the data are entirely treelike, this approach should, in theory, discover the single appropriate tree. Where the data are not compatible with a tree structure, this approach should find the best tree, along with measures indicating the extent to which that tree is compatible with the data. Readers should be cautioned in advance that the tree generated by this method is thus only valid when interpreted statistically in terms of its fit with the data.

In contrast to the split decomposition methodology, which a computer can perform very quickly, finding the best tree is a computationally difficult problem. Even with just twelve languages, there are 24.3 billion possible trees, so it is not feasible to assess the fit of every possible tree with today's technology. Instead, we employ a Markov Chain Monte Carlo (MCMC) method to search through the probability space of possible trees. First applied to linguistic data by Gray and Atkinson (2003), this method relies on Bayesian statistical techniques. At each iteration, MCMC compares the current tree to other probable trees (most of which are quite similar to the current best, but including one random tree), and if a better tree is found, it becomes the current tree. The process is run iteratively until the probabilities converge.

We implemented MCMC on the Alor-Pantar lexical dataset with several different models, using both MrBayes 3.2.1 (Ronquist and Huelsenbeck, 2003) and BEAST 1.7.2 (Drummond et al., 2012). We ran each model for at least 10 million iterations with a sample rate of 1000 and a burn-in of 25 percent. We did four runs on the same dataset for each model, and each converged after approximately 1.5 million iterations. The best performing model (i.e., that with the highest likelihood) was the relaxed Dollo model implemented in BEAST.[7] This model is particularly appropriate to linguistic data since it assumes that

[7] A combined gamma covarion model performed nearly as well as the Dollo model and

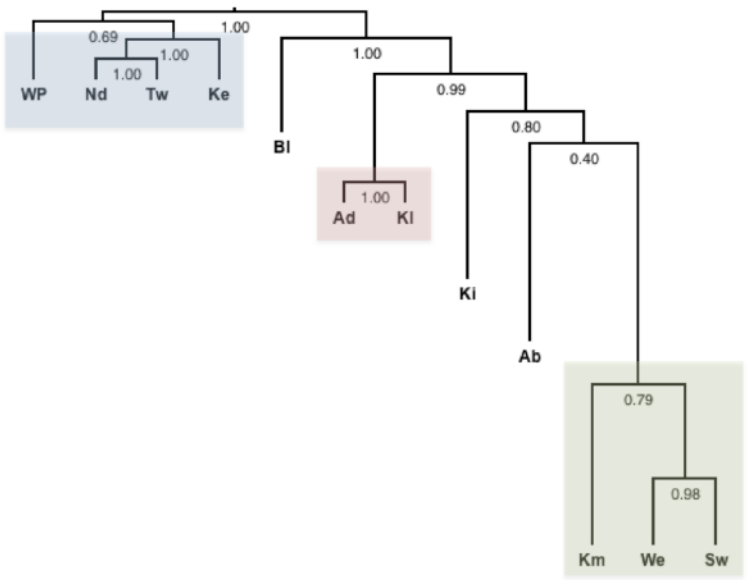

Figure 10. Bayesian MCMC consensus tree for lexical data (relaxed Dollo model)

innovations may arise only once, but may be lost multiple times independently (Pagel, 2009). The majority consensus tree for this model is shown in Fig. 10, with the pAP node used as an outgroup (not shown) to root the tree. The clade credibility values listed below each node indicate the percentage of sampled trees which are compatible with that node. Most of these values are either at or near one hundred percent, indicating that this consensus tree is compatible with almost all of the trees sampled in the analysis.

Before drawing any conclusions from this tree, a word of caution is in order. The fact that the Bayesian tree has high clade credibility values should not be interpreted as evidence that it is somehow the "right" or "correct" tree. Rather, the conclusion to be drawn is that additional searching is unlikely to reveal a tree which better fits our data. With this caution, we can compare the Bayesian tree to the split graph (Fig. 5). To a large extent the groupings in the Bayesian tree are compatible with those in the split graph. First, Sawila (Sw) and Wersing (We) are shown to be closely related, a grouping which was also present in the classification based on the comparative method (Fig. 4) and the split graph based on phono-

yielded similar results. In contrast to Dunn et al. (2011) we found a strict unidirectional model actually performed slightly better than the gamma covarion model (though still not as well as the Dollo model).

logical innovations (Fig. 9). Second, there is a Pantar grouping of Kaera (Ke), Teiwa (Tw), Nedebang (Nd), and Western Pantar (WP). Third, the position of Blagar at the highest node coordinate to the Alor languages is consistent with its position in the split graph—although, as noted above, this differs significantly from its position in the tree based on the traditional application of the comparative method (Fig. 4). On the other hand, there are also some incompatibilities between the Bayesian tree and the split graph. For example, in the Bayesian tree Adang (Ad) and Klon (Kl) are shown forming a group without Kui (Ki), contra both the splits graph and the tree calculated using the comparative method.

5.1. *Comparison to Lexicostatistics*

Given that we apply mathematical models to lexical data, we feel it is necessary to distinguish our work from that of early lexicostatistical methods (Swadesh, 1950). While it is clear that we owe an intellectual debt to these early methods, we would like to emphasize the differences in methodology. First, many early applications of lexicostatistics relied on subjective similarity judgments to identify lexical lookalikes. Indeed, this is the case with the Alor-Pantar data considered by Stokhof (1975). In our study, we instead assign cognacy based on prior application of the comparative method. Second, where the proto-forms are known, they have been included and coded as cognate with all the forms that are regular reflexes of that proto-form, thus enabling us to distinguish between innovation and retention, a crucial component of the comparative method. Third, our methods consider sets of cognates, as opposed to the pairwise comparisons of traditional lexicostatistics. That is, lexicostatistics misses out on the patterns in the data by simply comparing raw percentages of similarity (or sometimes cognacy) between pairs of languages. In contrast, the Bayesian methodology considers all the languages at once and picks out the tree that is most compatible with the entire dataset.

These theoretical differences between our computational methods and traditional lexicostatistics are also reflected in practice. To show this, we applied the lexicostatistical method to the same lexical dataset used for the split decomposition and Bayesian analyses above. Specifically, we employed the technique outlined in an introductory historical linguistics textbook (Crowley, 1997).[8] The results (Fig. 11) are strikingly different from those produced by the other com-

[8] Note that the method outlined by Crowley proposes grouping languages that have similar cognacy percentages, thus yielding a tree that is not binary. A Neighbor Joining algorithm in SplitsTree (Huson and Bryant, 2006) applied to the same dataset produced a very similar tree, with the positions of Blagar (Bl) and Western Pantar (WP) reversed and Klon (Kl) and Kui (Ki) slightly closer to each other than to Adang (Ad).

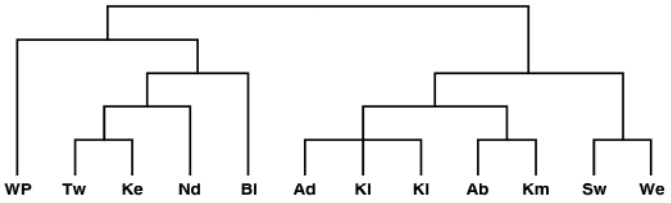

Figure 11. Tree generated using traditional lexicostatistics

putational methods. First, the lexicostatistical tree groups Blagar (Bl) within the Pantar languages (Nd, Tw, Ke, WP), whereas both computational methods group Blagar with the Alor languages. Second, the lexicostatistical tree groups Abui (Ab) and Kamang (Km) together, distinct from Sawila (Sw) and Wersing (We). This contrasts with the Bayesian consensus tree, which groups Kamang but not Abui closely with Sawila and Wersing. Third, the lexicostatistical tree groups Adang (Ad), Klon (Kl), and Kui (Ki), while in the Bayesian tree only the first two of these subgroup together. Further inspection will show that this tree is different in many more respects from the tree produced by the Bayesian methods.

We emphasize that we are not advocating traditional lexicostatistics as an alternative subgrouping methodology here. In particular, we assign no credibility to the tree in Fig. 11. We show it here merely to demonstrate that our methods are not only theoretically robust, but also yield results which are decidedly different from those derived using lexicostatistics.

6. Discussion

We draw three important conclusions from our analyses. First, the phylogenetic methods considered here reveal new information regarding the internal structure of the Alor-Pantar family. In contrast to the prevailing *Ethnologue* classification, subgrouping based on phylogenetic methods does not group Blagar with the Pantar languages. This is compatible with results from the traditional comparative method, with Blagar embedded deep within an Alor subgroup. However, in contrast to our results, the tree derived from the comparative method fails to delineate a Pantar subgroup; rather, the Pantar languages represent four primary groups of Proto-Alor-Pantar (cf. Fig. 4). Whereas the tree based on the comparative method points to an original settlement on Pantar, the tree based on computational methods is compatible with a social scenario in which the original speakers of proto-Alor-Pantar first entered the archipelago in the area of the Pantar Straits where Blagar is spoken today. The significant embedding within the remaining Alor languages (and to a lesser degree within Pantar) sug-

gests migration and settlement outwards from the Pantar Straits. This was not apparent in earlier studies but is confirmed here by both the split decomposition and Bayesian models. This result points to the need for further descriptive work on Blagar, particularly in the context of the Timor languages, which are the subject of comparative work currently in progress (Schapper et al., 2012).

A second important conclusion is that computational phylogenetic tools can be extremely powerful when combined with the traditional comparative method. Far from being a competing methodology, ~~the~~ phylogenetic methods complement the comparative method by illuminating precisely the area where the comparative method sheds least light. Even when the comparative method works well (as we believe is the case in Alor-Pantar), the phonological innovations identified by the method rarely give rise to clear, uncomplicated trees. The task of subgrouping is often complicated by wavelike patterns of borrowing and lexical innovations. By using knowledge of phonological innovations to identify non-cognate lexical items, we can clearly identify subgroups based on individual lexical items. The computational tools then permit us to infer aggregate information about subgroups, even where different lexical items yield different signals. Here the crucial insight is recognizing that there is no one correct tree representing a single history, but rather multiple overlapping trees representing both vertical and lateral transfer events (Hall, 2008). Underlying these methods is the strong foundation of the comparative method, which allows us to identify subgroups even when the phonological innovations themselves do not clearly delineate such subgroups.

Finally, our work demonstrates that computational phylogenetic tools can be extremely effective on a local level. Most previous approaches which employ the methodologies discussed here have focused on the world's major language families, such as Indo-European (Gray and Atkinson, 2003) or Austronesian (Gray and Jordan, 2000). To a certain extent this large-scale focus is reasonable, given the relatively new nature of these tools. In order to test and evaluate new tools, it is only natural to begin with larger-scale problems which have already received considerable attention within the field of historical linguistics. But as these tools mature, we see increasing applications of computational phylogenetic methods to the smaller-scale bread-and-butter problems of language classification and linguistic prehistory. Already we are beginning to see applications of these tools to the study of internal classification in families once thought to be resistant to traditional methods (e.g., Sicoli and Holton, 2012 on Athabaskan). No longer an exotic tool, phylogenetic methods have come into their own as a standard tool in historical linguistics.

Acknowledgements

Field work was supported by grants from the Netherlands Organization for Scientific Research, the UK Arts and Humanities Council, and the US National Science Foundation (NSF-SBE 0936887), under the aegis of the European Science Foundation EuroBABEL programme. The authors are indebted to their colleagues in the EuroBABEL Alor-Pantar project for generously sharing their data and analyses, and for providing feedback on early versions of this paper. The authors also wish to thank numerous colleagues in Alor and Pantar who assisted with data collection. Finally, this paper is greatly improved thanks to the comments of four anonymous reviewers, who of course are not responsible for any remaining errors of fact or interpretation.

References

Anonymous. 1914. De eilanden Alor en Pantar, Residentie Timor en Onderhoorigheden. *Tijdschrift van het Koninklijk Nederlandsch Aardrijkskundig Genootschap* 31: 70–102.

Bandelt, Hans-Jürgen and Andreas Dress. 1992. Split decomposition: A new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution* 1: 242–252.

Campbell, Lyle and William J. Poser. 2008. *Language Classification: History and Method*. Cambridge: Cambridge University Press.

Crowley, Terry. 1997. *An Introduction to Historical Linguistics*. 3rd ed. Oxford: Oxford University Press.

Donohue, Mark. 1996. Inverse in Tanglapui. *Language and Linguistics in Melanesia* 27: 101–118.

Donohue, Mark. 2007. The Papuan language of Tambora. *Oceanic Linguistics* 46: 520–537.

Donohue, Mark, Simon Musgrave, Bronwen Whiting, and Søren Wichmann. 2011. Typological feature analysis models linguistic geography. *Language* 87: 369–383.

Drummond, Alexei J., Marc A. Suchard, Dong Xie, and Andrew Rambaut. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* 29: 1969–1973.

Dunn, Michael, Niclas Burenhult, Nicole Kruspe, Sylvia Tufvesson, and Neele Becker. 2011. Aslian linguistic prehistory. *Diachronica* 28: 291–323.

Dunn, Michael, Stephen C. Levinson, Eva Linström, Ger Reesink, and Angela Terrill. 2008. Structural phylogeny in historical linguistics: Methodological explorations applied in island Melanesia. *Language* 84: 710–759.

Gordon, Raymond G., Jr. 2005. *Ethnologue: Languages of the World*. 15th ed. Dallas: S.I.L. International.

Gray, Russell and Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426: 435–439.

Gray, Russell, David Bryant, and Simon Greenhill. 2010. On the shape and fabric of human history. *Philosophical Transactions of the Royal Society* 365: 3923–3933.

Gray, Russell and Fiona Jordan. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature* 405: 1052–1055.

Hajek, John. 2010. Towards a phonological overview of the vowel and consonant systems of East Nusantara. In Michael Ewing and Marian Klamer (eds.), *Typological and Areal Analyses: Contributions from East Nusantara*, 25–46. Canberra: Pacific Linguistics.

Hall, Barry. 2008. *Phylogenetic Trees Made Easy: A How-to Manual*. 3rd ed. Sunderland, MA: Sinauer.

Holland, Barbara, Katharina Huber, Andreas Dress, and Vincent Moulton. 2002. δ Plots: A tool for analyzing phylogenetic distance data. *Molecular Biology and Evolution* 19: 2051–2059.

Holton, Gary, Marian Klamer, František Kratochvíl, Laura Robinson, and Antoinette Schapper. 2012. The historical relation of the Papuan languages of Alor and Pantar. *Oceanic Linguistics* 51: 87–122.

Huson, Daniel H. and David Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23: 254–267.

Huson, Daniel H. and Mike Steel. 2004. Phylogenetic trees based on gene content. *Bioinformatics* 20: 2044–2049.

Krauss, Michael E. 1973. ~~Eskimo-Aleut~~. In Thomas A. Sebeok (ed.), *Linguistics in North America*, ~~796–902~~. The Hague: Mouton.

Lewis, M. Paul. 2009. *Ethnologue: Languages of the World*. 16th ed. Dallas: S.I.L. International.

Nicolspeyer, Martha Margaretha. 1940. *De Sociaale Structuur van een Aloreesche Bevolkingsgroep*. Rijswijk: Kramers.

Pagel, Mark. 2009. Human language as a culturally transmitted replicator. *Nature Reviews Genetics* 10: 405–415.

Robinson, Laura and Gary Holton. To appear. Reassessing the wider genetic affiliations of the Timor-Alor-Pantar languages. *Language and Linguistics in Melanesia*.

Ronquist, Fredrik and John P. Huelsenbeck. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19: 1572–1574.

Schapper, Antoinette, Aone van Engelenhoven, and Juliette Huber. 2012. The historical relations of the Papuan languages of Timor and Kisar. Paper presented at the Conference on History, Contact and Classification of Papuan Languages, Amsterdam, February 2012.

Sicoli, Mark and Gary Holton. 2012. Applying computational phylogenetic methods in evaluation of the Dene-Yeniseian hypothesis. Paper presented at the Alaska Anthropological Association, Seattle, WA, March 2012.

Sidwell, Paul. 2009. How many branches in a tree? Cua and East (North) Bahnaric. In Bethwyn Evans (ed.), *Discovering History through Language. Papers in Honour of Malcom Ross*, 193–204. Canberra: Pacific Linguistics.

Stokhof, W.A.L. 1975. *Preliminary Notes on the Alor and Pantar Languages (East Indonesia)*. Canberra: Australian National University.

Swadesh, Morris. 1950. Salish internal relationships. *International Journal of American Linguistics* 16: 157–167.

Wichmann, Søren, Eric W. Holman, Taraka Rama, and Robert S. Walker. 2011. Correlates of reticulation in linguistic phylogenies. *Language Dynamics and Change* 1: 205–240.

Appendices

Five datasets relevant to this paper can be found online. These datasets derive from lexical data collected by the authors and their colleagues and first described in Holton et al. (2012). Reconstructions are those in Holton et al. (2012), augmented by additional analysis based on the work of the current authors. See http://dummy.dummy. ■ please add correct url

- AP_lexicon.txt is a tab-delimited UTF-8 text file containing the entire set of 400 lexical items in each of the twelve Alor-Pantar languages referred to in this paper, together with a reconstructed proto-Alor-Pantar form, where known.
- AP_lexicon_coded.txt is a tab-delimited UTF-8 text file containing 331 lexical items from AP_lexicon.txt, each coded into cognacy classes. The cognate classes are represented by numerals immediately below each reflex. Items not assigned to cognacy classes are indicated with a hyphen. The number of coded lexical items is significantly smaller than the total dataset because obvious recent borrowings have not been coded.
- AP_splits.nex is the NEXUS-formatted text file used to generate the split graph shown in Fig. 5. The character matrix is that obtained by converting the multistate codings in AP_lexicon_coded.txt to binary characters.
- AP_beast.xml is an XML text file formatted using BEAUTi 1.7.2 for analysis in BEAST 1.7.2. The content of the character matrix is identical to that in AP_splits.nex. This file was used to generate the trees of which Fig. 10 is a majority-rules consensus.
- delta_scores.txt is a tab-delimited UTF-8 text file containing a matrix of delta scores for each language by semantic category. This matrix was used to generate the principal components in Fig. 8.