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

Laura C. Robinson
Gary Holton

University of Alaska Fairbanks

October 15, 2011
Overview

• Background
  – previous comparative work
  – attempts at subgrouping

• Computational Methods
  – Phylogenetic network model
  – Bayesian tree model

• Conclusions
Background
Papuan in a sea of Austronesian
The data

• Since 2003, much new data has been collected on the AP languages
• The reconstruction via the comparative method has focused on 12 of the approximately 20 languages
• wide geographical sample
History of classification

- Capell (1943) (Josselin de Jong 1937, Nicolspeyer 1940) recognizes non-Austronesian character of AP languages
- Stokhof (1975) survey provides first clear comparative lexical data (117 item lists for 12 varieties)
- Holton, Klamer, Kratochvíl, Robinson, and Schapper (to appear) reconstruct pAP drawing on extensive field work from previous decade
Subgrouping: Ethnologue 15 (2005)
Subgrouping: Ethnologue 16 (2009)
An empirically based approach to subgrouping
### The comparative method

<table>
<thead>
<tr>
<th>pAP</th>
<th>Env</th>
<th>Tw</th>
<th>Nd</th>
<th>Ke</th>
<th>WP</th>
<th>Bl</th>
<th>Ad</th>
<th>Kl</th>
<th>Ki</th>
<th>Ab</th>
<th>Km</th>
<th>Sw</th>
<th>We</th>
</tr>
</thead>
<tbody>
<tr>
<td>*b</td>
<td>#_</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>f</td>
<td>p</td>
<td>p</td>
</tr>
<tr>
<td>*b</td>
<td>*b</td>
<td>f/v</td>
<td>f/v</td>
<td>b</td>
<td>bb</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>f</td>
<td>p</td>
<td>p</td>
</tr>
<tr>
<td>*b</td>
<td>*b</td>
<td>f/v</td>
<td>f/v</td>
<td>b</td>
<td>p</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>Ø</td>
<td>p</td>
<td>p</td>
<td></td>
</tr>
<tr>
<td>*d</td>
<td>*d</td>
<td>d</td>
<td>d</td>
<td>d</td>
<td>d</td>
<td>d</td>
<td>d</td>
<td>d</td>
<td>r</td>
<td>t</td>
<td>d</td>
<td>d</td>
<td></td>
</tr>
<tr>
<td>*d</td>
<td>*d</td>
<td>d</td>
<td>d</td>
<td>dd</td>
<td>d</td>
<td>d</td>
<td>d</td>
<td>d</td>
<td>r</td>
<td>t</td>
<td>d</td>
<td>d</td>
<td></td>
</tr>
<tr>
<td>*d</td>
<td>*d</td>
<td>r</td>
<td>r</td>
<td>d</td>
<td>r</td>
<td>d</td>
<td>d</td>
<td>d</td>
<td>r</td>
<td>t</td>
<td>d</td>
<td>d</td>
<td></td>
</tr>
<tr>
<td>*g</td>
<td>*g</td>
<td>g</td>
<td>g</td>
<td>g</td>
<td>g</td>
<td>?</td>
<td>?</td>
<td>g</td>
<td>g</td>
<td>h</td>
<td>g</td>
<td>g</td>
<td></td>
</tr>
<tr>
<td>*g</td>
<td>*g</td>
<td>h</td>
<td>x</td>
<td>g</td>
<td>gg</td>
<td>Ø</td>
<td>?</td>
<td>g</td>
<td>g</td>
<td>h</td>
<td>Ø</td>
<td>y</td>
<td>l</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gloss</th>
<th>pAP</th>
<th>Tw</th>
<th>Nd</th>
<th>Ke</th>
<th>WP</th>
<th>Bl</th>
<th>Ad</th>
<th>Kl</th>
<th>Ki</th>
<th>Ab</th>
<th>Km</th>
<th>Sw</th>
<th>We</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘pig’</td>
<td>*b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>f</td>
<td>p</td>
<td>p</td>
</tr>
<tr>
<td>*bay</td>
<td>bai</td>
<td>bei</td>
<td>bei</td>
<td>bai</td>
<td>be</td>
<td>bi</td>
<td>be:?</td>
<td>bei</td>
<td>fe</td>
<td>pe:</td>
<td>pi</td>
<td>pei</td>
<td></td>
</tr>
<tr>
<td>*bui</td>
<td>bui</td>
<td>buya</td>
<td>bai</td>
<td>bu</td>
<td>bu</td>
<td>bu:</td>
<td>boih</td>
<td>bai</td>
<td>fu</td>
<td>pu</td>
<td>pui</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Phonological Innovations

<table>
<thead>
<tr>
<th>Innovation</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>*h &gt; Ø</td>
<td>Everywhere but Tw and WP</td>
</tr>
<tr>
<td>*q &gt; k</td>
<td>Everywhere but Tw, Nd, Ke (Ad ? &lt; k &lt; *q)</td>
</tr>
<tr>
<td>*n &gt; ŋ / _#</td>
<td>Everywhere but Tw, Kl, Ki</td>
</tr>
<tr>
<td>*l &gt; Ø / _#</td>
<td>Nd, WP, Ab</td>
</tr>
<tr>
<td>*r &gt; Ø / _#</td>
<td>Tw, Ke, WP</td>
</tr>
<tr>
<td>*b &gt; f</td>
<td>Tw, Nd, Ab (in Tw and Nd only non-initially)</td>
</tr>
<tr>
<td>*l &gt; y / _#</td>
<td>Tw, Ke, Ad, Km</td>
</tr>
<tr>
<td>*r &gt; l / V_V</td>
<td>Nd, WP, Ad, Km</td>
</tr>
<tr>
<td>*m &gt; ŋ / _#</td>
<td>WP, Bl, Ad</td>
</tr>
<tr>
<td>*g &gt; ?</td>
<td>Bl, Ad</td>
</tr>
<tr>
<td>*k &gt; Ø / _#</td>
<td>Bl, Ad</td>
</tr>
<tr>
<td>*r &gt; y / _#</td>
<td>Bl, Ki, Ab</td>
</tr>
<tr>
<td>*s &gt; h</td>
<td>Bl, Ad, Kl</td>
</tr>
<tr>
<td>*d &gt; r</td>
<td>Ab, Ki (in Ki only finally)</td>
</tr>
<tr>
<td>*s &gt; t</td>
<td>Ab, Sw, We</td>
</tr>
<tr>
<td>*b &gt; p</td>
<td>Km, Sw, We</td>
</tr>
<tr>
<td>*y &gt; Ø / _#</td>
<td>Km, Sw</td>
</tr>
</tbody>
</table>
Subgrouping based on phonological innovations
Cross-cutting innovations

• *s > h in Bl, Ad, Kl
  but *m > η in WP, Bl, Ad

• *b > p in Km, Sw, We
  but *s > t in Ab, Sw, We

• *l > Ø / _# in WP, Nd, Ab
  but *r > Ø / _# in WP, Tw, Ke
Lexical innovations and subgrouping

- phonological innovations do not yield neat subgroups,
- but they *do* allow us to identify inherited vs. borrowed forms (as long as the borrowings are relatively recent)
- we can then use the lexicon as a tool for subgrouping
## Detecting lexical innovations

<table>
<thead>
<tr>
<th></th>
<th>‘sharp’</th>
<th>medial *g</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teiwa</td>
<td>maḥan</td>
<td>ḫ</td>
</tr>
<tr>
<td>Nedebang</td>
<td>maxaŋ</td>
<td>x</td>
</tr>
<tr>
<td>Kaera</td>
<td>magaŋ</td>
<td>g</td>
</tr>
<tr>
<td>W Pantar</td>
<td>maggaŋ</td>
<td>gg</td>
</tr>
<tr>
<td>Blagar</td>
<td>manŋ</td>
<td>Ø</td>
</tr>
<tr>
<td>Klon</td>
<td>manŋ</td>
<td>g</td>
</tr>
<tr>
<td>Kui</td>
<td>maŋan</td>
<td>g</td>
</tr>
</tbody>
</table>

Klon and Kui cannot be cognate with the others. They may be borrowed from Blagar.
Detecting lexical innovations

<table>
<thead>
<tr>
<th></th>
<th>‘small’</th>
<th>*d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kui</td>
<td>kadin</td>
<td>d</td>
</tr>
<tr>
<td>Abui</td>
<td>kidiŋ</td>
<td>r</td>
</tr>
<tr>
<td>Kamang</td>
<td>kidiŋ</td>
<td>t</td>
</tr>
</tbody>
</table>

Cannot be cognate; perhaps borrowed from Kui *kadin*
Lexical innovations and subgrouping

• Lexical innovations can be useful for subgrouping
• but we have 400 items and 12 languages
• Combing through 400 lexical items in 12 languages (plus pAP) to look for patterns is difficult
• Lexical isoglosses are not obviously bundled
• Hard to recognize tree-like signals by eye
Phylogenetic network model
Phylogenetic network model

Insight from biology:
Recombination may yield two different histories.

“It is essential to understand that the ‘right tree’ doesn’t exist.”
(Hall 2008)
Implementing the network model

• 351 lexical items coded for cognacy classes
• Code pAP as well so that retentions can be readily identified
• Code cognate classes as numeric states to yield a 13x351 matrix (13x2542 binary!)
• Use SplitsTree (Huson & Bryant 2006) to do split decomposition with NeighborNet algorithm
AP network

\[ \delta = 0.35 \]
‘crocodile’
‘wind’

Note: Network changes slightly because pAP data is excluded
‘rain’

Note: Network changes slightly because pAP data is excluded.
AP network

[Diagram of AP network with labels: We, Sw, WP, Km, Ab, Ki, Kl, Ad, Bl, Nd, Tw, Ke]
Bayesian tree model
Bayesian tree model

• Even though the ‘right tree’ may not exist, it may be possible to find the ‘best’ tree in some sense
• Procedure: take an informed walk through space of possible trees looking for the tree with the greatest probability
Implementing Bayesian tree model

• take the same character matrix
• use MrBayes (Ronquist & Huelsenback 2003) to generate increasingly probable trees from a set of 24.3 billion possible trees
• stop looking when probabilities converge (10 m iterations)
Bayesian tree model
Lexicostatistics vs. Bayesian

Lexicostatistics

Bayesian tree
Why different from lexicostatistics?

• where pAP is known, it is included
  – thus we are looking at shared *innovation* not retention

• Even with the same cognate coding, the Bayesian and Network models are more complex than lexicostatistical models

• items are coded for *cognacy* not similarity
Network vs. Bayesian tree

Network

Bayesian tree

[Diagram showing network and Bayesian tree structures with various nodes and connections labeled with letters like We, Sw, WP, Km, Ab, Tw, Ke, Ki, Nd, pAP, Bl, Ad, etc., with some branches indicating probabilities like 0.99 or 1.00.]
Phonology vs. Lexicon

phonological innovations

lexicon

43 characters

351 characters
Variation across the lexicon

Lexical Items by Semantic Category
Teasing out contribution of individual semantic categories

• Not sufficiently independent to do a factor analysis ($p = .45$)

• But we can extract principal components
Teasing out contribution of individual semantic categories
Teasing out contribution of individual semantic categories
First two components
Network based on time/location
Conclusions
Defining subgroups based on lexical innovations

• We use the comparative method to identify phonological innovations
• We use phonological innovations to identify lexical innovations
• We use lexical innovations to identify subgroups
• This works even when the phonological innovations don’t identify subgroups
Summary

• Computational phylogenetic tools can be extremely powerful, especially when combined with the Comparative Method

• Prior application of Comparative Method yielded robustly supported phonological innovations

• However, innovations do not delineate neat subgroups

• Crucially, this is not a failure of the Comparative Method
Beyond the Comparative Method

- Computational techniques should not be viewed as a replacement for or alternative to the Comparative Method.
- Rather, in the application discussed here computational approaches supplement the Comparative Method where the latter fails to yield subgrouping information.
Why computational methods succeed where the Comparative Method does not

• AP phonological innovations
  – are extremely typologically common
  – thus they may be similar due to chance

• But, lexical innovations almost certainly not due to chance
  – We use Comparative Method as a tool for coding the lexical data based cognacy
  – with nearly 400 lexical items, the data are so complex that you can’t sort out subgroups by hand unless they’re perfect (i.e., no borrowing).
  – Too easy to borrow from a language with same phonological history (and thus borrowings are undetected)
Remaining issues

• Cognate classes are only as good as our pAP reconstruction
  – still, we believe an approach to coding which makes use of the comparative method is better than one which relies on similarity judgments

• Coding remains an issue
  – how much semantic leeway should we allow in assessing cognacy
Remaining issues

• How to assess cognacy for items which do not reflect pAP
  – e.g., *b > p in Kamang, Wersing
  – but, Kamang bongko, Wersing bongkau ‘cassava’ both have /b/
  – these were still counted as cognate with each other (and not inherited from pAP)
    • BUT might be loans from AP or elsewhere
Future work

• Include data from Timor languages
  – but we need to do comparative method first
• Include data from mainland New Guinea languages (e.g., Bomberai)
  – in progress (Amsterdam Feb 2012)
Selected References

• Holton, G., M.A.F. Klamer, F. Kratochvíl, L.C. Robinson, & A. Schapper. To Appear. The historical relation of the Papuan languages of Alor and Pantar. OL.
Acknowledgements

US National Science Foundation grant NSF-SBE 0936887