A Bayesian Phylogenetic Internal Classification of the Tupí-Guaraní Family

Lev Michael, Natalia Chousou-Polydouri, Zachary O’Hagan, Keith Bartolomei, Diamantis Sellis, Emily Clem, and Erin Donnelly

University of California, Berkeley

~

Linguistic Society of America
Annual Meeting
Portland, Oregon
January 10, 2015
The adoption of computational phylogenetic methods originally developed in biology has generated considerable high-profile work in historical linguistics in recent years:

- **Indo-European**: Bouckaert et al. (2012); Forster and Toth (2003); Gray and Atkinson (2003); Nakhleh et al. (2005); Ringe et al. (2002); Warnow et al. (2004)
- **Austronesian**: Gray et al. (2009); Greenhill and Gray (2005, 2009); Greenhill et al. (2010)
- **Pama-Nyungan**: Bowern and Atkinson (2012)

This research has focused principally on the application of phylogenetic methods to lexical data.
• The successes in applying phylogenetic methods to historical linguistics is to be expected in certain respects
  • Biological phylogenetics is based on a model of evolution that is compatible with linguists’ understanding of diachronic change
  • Both biological and linguistic evolution involve descent with modification from a common ancestor, which gives rise to primarily tree-like evolutionary histories
Introduction III

• At the same time, valid application of phylogenetic methods to linguistic data that is both
  • an accurate implementation of the ideas of the Comparative Method
  • and does not violate the mathematical assumptions behind the computational methods

...is not a trivial matter

• In this talk, we **first** examine the current standard application of phylogenetic methods to comparative lexical data (Gray and Atkinson 2003), which we dub the ‘G&A method’ and argue that it:
  1. does not implement linguists’ understanding of cognacy;
  2. introduces problematic mathematical artifacts (character non-independence) due to coding implementations

• And **second**, present and compare an alternative method, quasi-cognate coding (‘QC’), that we argue:
  1. more faithfully implements linguists’ understanding of cognacy;
  2. minimizes coding-induced character non-independence
• We compare the results of these two methods when applied to a lexical dataset of Tupí-Guaraní languages
• The two methods yield different results:
  • The QC coding reveals higher-level structure that the G&A coding results do not identify
  • The QC coding accords better with previous classifications
Applying Computational Phylogenetics to Linguistic Data

- Phylogenetic methods share basic principles with the Comparative Method, and have become increasingly sophisticated via:
  - evolutionary models
  - Bayesian inference methods
  - computational algorithms

- The application of these tools and methods to linguistic data
  - is not a mechanical procedure;
  - and requires careful thought about the nature of the data and the phylogenetic characters to be extracted from that data

- Here we focus on the impact of these early methodological decisions, prior to phylogenetic analysis per se, on the phylogenetic results:
  - lexical data collection
  - cognate set construction and character coding
Gray and Atkinson (2003) introduced the current standard for applying computational phylogenetic tools to lexical data:

- given a set of meanings (e.g., Swadesh list), a single form is selected
- for each meaning, forms are grouped into \( n \) “cognate” sets, resulting in a \( n \)-state (multistate) character per meaning
  - The characters are thus members of ‘form-meaning’ sets in which forms are cognate and have the same meaning
  - Such sets are not cognate sets, since forms that may have undergone semantic shift are not members of the same ‘form-meaning’ set
- each multistate character state is recoded as a binary presence-absence character
Potential Problems wth G&A Coding

- Loss of ability to capture synapomorphies: features inherited from a common ancestor
  - When a given cognate set is split into multiple form-meaning sets, the fact that the form-meaning sets are related to each other is lost
- Introduction of homoplasies: shared features **not** inherited from a common ancestor
  - Common semantic shifts may occur independently (e.g., ‘dark’ → ‘night’)
  - Thus form-meaning sets based on these meanings conflate multiple origin events (unlike true cognate sets)
- Binary recoding of originally multi-state characters introduces – for mathematical reasons – non-independence between the resulting binary characters
  - This violates a key assumption of phylogenetic algorithms
Quasi-Cognate Coding Method I

- The potential problems associated with the G&A method led us to develop a method – the quasi-cognate (QC) method – that more closely hews to the assumptions of the Comparative Method.
- The quasi-cognate method is characterized by the following:
  - Characters are members of true cognate sets (irrespective of meaning).
  - Characters are binary (a language has a word that is a member of a given cognate set, or it does not).
Quasi-Cognate Coding Method II

- **Data Collection, Round 1:**
  1. Given a set of meanings, collect for each language *all* forms with the meaning in question, *as well those with similar meanings*
  2. Construct cognate sets that include items that have undergone semantic shift

- **Data Collection, Round 2:** in case of apparent absences in cognate sets, search for cognates with the expected form (given deducible sound correspondences) for the language in question

- *(The resulting cognate sets are quasi-independent, since we still assume, that in the absence of contrary evidence, if a language has a form expressing meaning A, the language lacks cognates for all other roots primarily associated with meaning A.)*
Zeroes & the G&A Method

- For purposes of computational analysis we numerical code our dataset was binary (1/0) presence/absence characters.
- ‘Zeroes’ (0s), indicating absence of a cognate for a particular language, play as significant of a role in the selection of an optimal tree as do ‘ones’ (1s), which indicate the presence of a cognate.
- It is thus important that a ‘0’ in the character table reflect – to as great a degree as possible – a true absence of a cognate in the language, and not merely a gap in documentation or data collection.
- However, in the G&A method, a language receives a ‘0’ for a given form-meaning set either if a cognate has undergone semantic shift, or if there is an empirical gap in the resource.
- Consequently, as normally implemented, this method does not distinguish documentation gaps from true absences.
Zeroes & the QC Coding Method

- We applied a more rigorous standard to ensure that a ‘0’ reflects a true absence
- A cognate was considered absent (coded as ‘0’) for a particular language if all the following conditions were met:
  1. No cognate was found when searching for roots with similar meanings, or for expected forms for the root;
  2. No cognate surfaced in compounds in our dataset;
  3. A non-cognate form was found that expressed the expected meaning
# Example of G&A vs. QC Coding

<table>
<thead>
<tr>
<th></th>
<th>OMG</th>
<th>KK</th>
<th>TPN</th>
<th>TPR</th>
<th>PAR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>G&amp;A (Multistate)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Woman</td>
<td>wajnu</td>
<td>wajna</td>
<td>kujã</td>
<td>koʃi</td>
<td>koʃo</td>
</tr>
<tr>
<td>Sister</td>
<td>kunia</td>
<td>kuña</td>
<td>ikœr</td>
<td>iket</td>
<td>ikœr</td>
</tr>
<tr>
<td><strong>G&amp;A (Binary Recoding)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WOMAN1 (*wajnu)</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>WOMAN2 (*kujã)</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SISTER1 (*kujã)</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SISTER2 (*ikœr)</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>QC (Binary)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WOMAN1 (*wajnu)</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>WOMAN2 (*kujã)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SISTER1 (*ikœr)</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Phylogenetic Methods (MrBayes3.2)

- We used an asymmetric binary model (a.k.a. restriction site model)
  - Different rates of gain and loss for cognates
  - Uniform prior for the cognate loss/gain ratio
- We allowed for different rates of evolution across cognate sets
  - Gamma distributed rates
  - Gamma shape parameter had a uniform prior distribution for (0,200)
- Phylogenetic Analysis with MrBayes3.2
  - Analysis conducted with four independent runs
  - 10 million generations each, sampled every 1,000 generations
The Tupí-Guaraní Family


- First phylogenetic exploration of Tupí: Galúcio et al. (2013)

![Figure 1: Tupí Classification (Drude 2011)]
Data Harvesting

- There are \( \sim 53 \) Tupí-Guaraní varieties spoken in Argentina, Bolivia, Brazil, Colombia, French Guiana, Paraguay, Peru
  - Degree of lexical documentation varies widely
- The lexical database developed for this project includes:
  - 596-item list of crosslinguistically and areally appropriate meanings in
  - 30 TG and 2 non-TG Tupí languages (Mawé and Awetí)
- Data was harvested by Keith Bartolomei, Natalia Chousou-Polydouri, Erin Donnelly, Lev Michael, Sérgio Meira, Zachary O’Hagan, Mike Roberts, and Vivian Wauters from:
  - dictionaries
  - phonological descriptions
  - grammatical descriptions
  - text collections
- Average coverage = 71%
### Lexical Coverage

<table>
<thead>
<tr>
<th>Language</th>
<th>Coverage</th>
<th>Language</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aché</td>
<td>85%</td>
<td>Ñandeva</td>
<td>20%</td>
</tr>
<tr>
<td>Anambé</td>
<td>31%</td>
<td>Omagua</td>
<td>89%</td>
</tr>
<tr>
<td>Araweté</td>
<td>55%</td>
<td>Parakanã</td>
<td>75%</td>
</tr>
<tr>
<td>Avá-Canoeiro</td>
<td>51%</td>
<td>Paraguayan Guarani</td>
<td>94%</td>
</tr>
<tr>
<td>Awetí</td>
<td>76%</td>
<td>Parintintin</td>
<td>85%</td>
</tr>
<tr>
<td>Chiriguano</td>
<td>80%</td>
<td>Pauserna</td>
<td>58%</td>
</tr>
<tr>
<td>Emerillon</td>
<td>77%</td>
<td>Siriono</td>
<td>82%</td>
</tr>
<tr>
<td>Guajá</td>
<td>45%</td>
<td>Tapiete</td>
<td>84%</td>
</tr>
<tr>
<td>Guarayu</td>
<td>86%</td>
<td>Tapirapé</td>
<td>69%</td>
</tr>
<tr>
<td>Ka’apor</td>
<td>83%</td>
<td>Tembé</td>
<td>98%</td>
</tr>
<tr>
<td>Kaiowá</td>
<td>39%</td>
<td>Tocantins Asuriní</td>
<td>83%</td>
</tr>
<tr>
<td>Kamaiurá</td>
<td>75%</td>
<td>Tupinambá</td>
<td>94%</td>
</tr>
<tr>
<td>Kayabí</td>
<td>59%</td>
<td>Wayampí</td>
<td>89%</td>
</tr>
<tr>
<td>Kokama</td>
<td>89%</td>
<td>Xetá</td>
<td>33%</td>
</tr>
<tr>
<td>Mawé</td>
<td>80%</td>
<td>Xingú Asuriní</td>
<td>50%</td>
</tr>
<tr>
<td>Mbyá</td>
<td>83%</td>
<td>Yuki</td>
<td>80%</td>
</tr>
</tbody>
</table>
Tupí-Guaraní Classification: QC Coding
Tupí-Guaraní Classification: G&A Coding

- Aweti
  - Paraguay-Guarani
    - Kaiowa
    - Mbya
    - Nandeva
  - Tapiete
    - Chiriguano
    - Xeta
  - Guarayu
    - Pauserna
  - Siriono
    - Omagua
    - Yuki
  - Tupinamba
    - Wayampi
      - Emerillon
    - Tapirape
      - Tocantins-Asurini
      - Parakana
      - Anambe
      - Arawete
      - Xingu-Asurini
    - AvaCanoeiro
      - Tembe
      - Kayabi
      - Parintintin
      - Kamaiura
      - Kaapor
      - Guaja
### Comparison of Classifications I

<table>
<thead>
<tr>
<th>QC</th>
<th>G&amp;A</th>
</tr>
</thead>
<tbody>
<tr>
<td>More higher-level articulation</td>
<td>Less higher-level articulation</td>
</tr>
<tr>
<td>More unique nodes</td>
<td>Fewer unique nodes</td>
</tr>
<tr>
<td>Kamaiurá is sister to ‘Nuclear-TG’</td>
<td>Kamaiurá is part of polytomy</td>
</tr>
<tr>
<td>Member of subgroup (Tupinambá, Avá-Canoeiro)</td>
<td>Merged with polytomy (Tupinambá, Avá-Canoeiro)</td>
</tr>
</tbody>
</table>

- G&A results can mostly be obtained from the QC results by eliminating higher-level structure, and merging the affected languages and subgroups into large polytomies.
- They differ in that the loss of higher-level structure in the G&A tree means that well supported subgroups in the QC coding disappear.
Comparison of Classifications II

• Which set of results is more plausible?
• We suggest that the classification that best captures the low-level subgroups recognized by TG specialists is preferred
  • Specialists are likely to have valid intuitions about low-level subgroups
  • There is little consensus regarding higher-level subgroups in the family
• How do the QC results compare with traditional classifications?
• Rodrigues and Cabral (2002): 8 subgroups of 44 TG varieties
  • Modification of the 8 subgroups of Rodrigues (1984/1985)
G&A & QC Tupí-Guaraní Classifications I
## Table 1: Subgroups Recovered by G&A and QC Codings

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>G&amp;A</th>
<th>QC</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>II</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>III</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>IV</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>V</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>VI</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>VII</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>VIII</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
Conclusions

• It is clear that G&A & QC coding produce significantly different results when applied to our TG lexical dataset

• We have argued in favor of quasi-cognate coding that it
  • better reflects linguists’ understandings of what constitutes suitable lexical phylogenetic characters;
  • does not violate the character independence assumption of phylogenetic methods;
  • better accords with traditional classifications of lower-level subgroups in the TG family

• It remains an open question to what degree the evident superiority of QC over G&A coding for the TG data set extends to comparable ones for other language families
Acknowledgements

• The following colleagues for generously sharing primary data:
  • Sebastian Drude (Awetí)
  • Sérgio Meira (Mawé, Tembé)
  • Françoise Rose (Emerillon)
  • Eva-Maria Rößler (Aché)
  • Rosa Vallejos (Kokama-Kokamilla)

• Tammy Stark for GIS assistance

• Noé Gasparini for access to Anambé and Yuki data

• And the following Berkeley TG group alumni:
  • Mike Roberts
  • Vivian Wauters

• NSF DEL Award #0966499

• UC Berkeley Social Science Matrix 2013-2014 Grant
References I


Galúcio, Ana Vilacy; Sérgio Meira; Sebastian Drude; Nilson Gabas Jr.; Denny Moore; Gessiane Picanço; Carmen Reis Rodrigues; and Luciana Storto. 2013. Genetic Relationship and Degree of Relatedness within the Tupi Linguistic Family: A Lexicostatistical and Phylogenetic Approach. *ms.*


References III


Warnow, Tandy; Steven N. Evans; Don Ringe; and Luay Nakhleh. 2004. Stochastic Models of Language Evolution and an Application to the Indo-European Family of Languages. *Technical Report, Department of Statistics, University of California, Berkeley.*
Lexicostatistics ≠ Phylogenetics

- Lexicostatistical Methods (e.g., NeighborNet, SplitsTree)
  - Lexicostatistical methods do not evaluate evolutionary trees
  - They instead compute a single number – e.g., % of shared cognates – for each pair of languages
  - Languages are then clustered on the basis of overall similarity, conflating shared innovations and shared retentions

- Phylogenetic Methods
  - All cognate sets are evaluated individually, and the specific information they bear for subgrouping is preserved
  - Thousands of trees are individually evaluated by optimizing all characters on each one
  - Only shared innovations are considered for subgrouping
  - As a result, phylogenetic methods are not fooled by shared retentions
## Homoplasy in G&A

<table>
<thead>
<tr>
<th></th>
<th>GYU</th>
<th>TPN</th>
<th>TPR</th>
<th>PAR</th>
<th>KAA</th>
</tr>
</thead>
<tbody>
<tr>
<td>stomach</td>
<td>?ie</td>
<td>iwe</td>
<td>...</td>
<td>aw</td>
<td>p+i+a</td>
</tr>
<tr>
<td>intestines</td>
<td>epi?i</td>
<td>iβiη</td>
<td>ie</td>
<td>ie</td>
<td>p+i+a</td>
</tr>
<tr>
<td>liver</td>
<td>p+i+a</td>
<td>p+i+a</td>
<td>p+i+a</td>
<td>p+i+a</td>
<td>p+i+a</td>
</tr>
</tbody>
</table>

### G&A (Multistate)

<table>
<thead>
<tr>
<th></th>
<th>GYU</th>
<th>TPN</th>
<th>TPR</th>
<th>PAR</th>
<th>KAA</th>
</tr>
</thead>
<tbody>
<tr>
<td>stomach</td>
<td>1</td>
<td>1</td>
<td>?</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>intestines</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>liver</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

### G&A (Binary)

<table>
<thead>
<tr>
<th></th>
<th>GYU</th>
<th>TPN</th>
<th>TPR</th>
<th>PAR</th>
<th>KAA</th>
</tr>
</thead>
<tbody>
<tr>
<td>STOMACH1 (*iwe)</td>
<td>1</td>
<td>1</td>
<td>?</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>STOMACH2 (*aβ)</td>
<td>0</td>
<td>0</td>
<td>?</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>STOMACH3 (*p+i+a)</td>
<td>0</td>
<td>0</td>
<td>?</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>INTESTINES1 (*epi?i)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>INTESTINES2 (*iβiη)</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>INTESTINES3 (*iwe)</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>INTESTINES4 (*p+i+a)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>LIVER1 (*p+i+a)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

### QC (Binary)

<table>
<thead>
<tr>
<th></th>
<th>GYU</th>
<th>TPN</th>
<th>TPR</th>
<th>PAR</th>
<th>KAA</th>
</tr>
</thead>
<tbody>
<tr>
<td>STOMACH1 (*aβ)</td>
<td>0</td>
<td>0</td>
<td>?</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>INTESTINES1 (*epi?i)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>INTESTINES2 (*iβiη)</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>INTESTINES3 (*iwe)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>LIVER1 (*p+i+a)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>