Southern Tupí-Guaraní Languages
Combining classification with areal linguistics

Noé Gasparini ¹  Swintha Danielsen ²  Natalia Chousou-Polydouri ¹
Zachary O’Hagan ³  Keith Bartolomei ³  Lev Michael ³

¹ University of Lyon, DDL
² University of Leipzig
³ University of California, Berkeley

The 18th Annual Workshop on American Indigenous Languages
May 8th 2015
Southern Tupí-Guaraní languages
Two subgroups (Rodrigues and Cabral 2012)
Three clades (Michael et al. to appear)
General idea: Phylogenetic methods can, via their ability to reconstruct ancestral states, be used to identify pools of possible loanword candidates.

Classifications

Phylogenetic classification

Looking for contact

Next steps
Previous classifications did often not include Yuki (Dietrich 1990; Rodrigues and Cabral 2012), Guarasu-Pauserna (Jensen 1998) and/or Jorá (Dietrich 2010).

<table>
<thead>
<tr>
<th>Quadro I: Nova constituição interna da família Tupí-Guaraní</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ramo I:</strong></td>
</tr>
<tr>
<td>• Guaraní Antigo</td>
</tr>
<tr>
<td>• Kaiwá (Kayová, Pãi), Ñandeva (Txiripá), Guaraní Paraguai</td>
</tr>
<tr>
<td>• Mbyá</td>
</tr>
<tr>
<td>• Xetá (Serra dos Dourados)</td>
</tr>
<tr>
<td>• Tapieté, Chiriguano (Ava), Izoceño (Chané)</td>
</tr>
<tr>
<td>• Guayakí (Aché)</td>
</tr>
<tr>
<td><strong>Ramo II:</strong></td>
</tr>
<tr>
<td>• Guarayo (Guarayú),</td>
</tr>
<tr>
<td>• Sirionó, Horá (Jorá)</td>
</tr>
<tr>
<td>• Yúki</td>
</tr>
</tbody>
</table>

Rodrigues and Cabral (2002:335)
Recent changes

- A lexical database in progress in Berkeley (Michael et al. to appear)
- Integration of all existing Jorá data collected by Danielsen and Gasparini (Submitted)
- New data from the field on Siriono
- Connections and collaborative works
Phylogenetic method

• A classification of Tupí-Guaraní based on lexical data and computational phylogenetic methods developed in evolutionary biology and more recently applied to linguistic phylogenies
  • See Nakhle et al. (2005); Gray et al. (2009); Greenhill et al. (2010); Bouckaert et al. (2012)
• This classification complements previous proposals using lexicostatistics, geographical proximity or sound change
Data Harvesting

- 31 TG and 2 non-TG Tupí languages (Mawé and Awetí)
- 595-item list of cross linguistically and areally appropriate meanings
- Organized into 4250 cognate sets
- Data were harvested by authors (Noé Gasparini, Swintha Danielsen, Natalia Chousou-Polydouri, Zachary O’Hagan, Keith Bartolomei, Lev Michael) and Sérgio Meira, Mike Roberts, Vivian Wauters, Erin Donnelly, from:
  - dictionaries, phonological descriptions, grammatical descriptions, text collections and field data
- Average coverage = 70%
### Lexical Coverage

<table>
<thead>
<tr>
<th>Language</th>
<th>Aché</th>
<th>85%</th>
<th>Guarayu</th>
<th>85%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kaiowá</td>
<td>39%</td>
<td></td>
<td>Guarasu</td>
<td>63%</td>
</tr>
<tr>
<td>Ñandeva</td>
<td>20%</td>
<td></td>
<td>Siriono</td>
<td>92%</td>
</tr>
<tr>
<td>Xetá</td>
<td>32%</td>
<td></td>
<td>Yuki</td>
<td>82%</td>
</tr>
<tr>
<td>Tapiete</td>
<td>84%</td>
<td></td>
<td>Jorá</td>
<td>21%</td>
</tr>
<tr>
<td>Chiriguano</td>
<td>80%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paraguayan Guarani</td>
<td>94%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mbyá</td>
<td>84%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Omagua</td>
<td>89%</td>
<td></td>
<td>Kokama</td>
<td>88%</td>
</tr>
<tr>
<td>Tupinambá</td>
<td>94%</td>
<td></td>
<td>Tembé</td>
<td>98%</td>
</tr>
<tr>
<td>Wayampí</td>
<td>89%</td>
<td></td>
<td>Emerillon</td>
<td>77%</td>
</tr>
<tr>
<td>Kayabí</td>
<td>59%</td>
<td></td>
<td>Parintintin</td>
<td>85%</td>
</tr>
<tr>
<td>Avá-Canoeiro</td>
<td>51%</td>
<td></td>
<td>Ka’ápor</td>
<td>83%</td>
</tr>
<tr>
<td>Guajá</td>
<td>45%</td>
<td></td>
<td>Tapirapé</td>
<td>69%</td>
</tr>
<tr>
<td>Tocantins Asurini</td>
<td>83%</td>
<td></td>
<td>Parakanã</td>
<td>75%</td>
</tr>
<tr>
<td>Anambé</td>
<td>31%</td>
<td></td>
<td>Araweté</td>
<td>55%</td>
</tr>
<tr>
<td>Xingú Asurini</td>
<td>51%</td>
<td></td>
<td>Kamaiurá</td>
<td>75%</td>
</tr>
<tr>
<td>Mawé</td>
<td>77%</td>
<td></td>
<td>Awetí</td>
<td>76%</td>
</tr>
</tbody>
</table>
Computational Phylogenetic Methods

• Phylogenetic characters: presence or absence of a cognate in each language, including words that have undergone semantic shift

• Model of evolution:
  • Potentially asymmetric model of gain and loss of cognates
  • Gamma distributed rate variation across cognate sets

• Phylogenetic analysis: Bayesian analysis using MrBayes3.2

• 4 independent runs of 10 million generations each

• Phylogenetic methods aim to find evolutionary trees that account for the distribution of set traits (= ‘characters’) across languages. Bayesian methods estimate the posterior probability of the tree topology and the model parameters given the data.
New phylogenetic classification, 2015
Rodrigues and Cabral (2002) and phylogenetic classification
Possible sources of loanwords
We performed maximum likelihood ancestral state reconstructions using the parameter estimates from the Bayesian analysis. For three different nodes in the tree: Guaranian, Guarayan and Sirionoan

- Generated lists of shared lexical innovations as a pool of potential borrowings.
  - Guaranian: 11 elements (3 unique)
  - Guarayan: 8 elements (3 unique)
  - Sirionoan: 61 elements (52 unique)

A clear split

Siriono, Yuki and Jorá show a substantial amount of differences.
<table>
<thead>
<tr>
<th>Language</th>
<th>Family</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baure</td>
<td>Arawakan</td>
<td>Danielsen database</td>
</tr>
<tr>
<td>Paunaka</td>
<td>Arawakan</td>
<td>Danielsen database</td>
</tr>
<tr>
<td>Bésiro</td>
<td>Macro-Gê</td>
<td>Courtesy of Pierric Sans</td>
</tr>
<tr>
<td>Ayoreo</td>
<td>Zamucoan</td>
<td>Courtesy of Luca Ciucci</td>
</tr>
<tr>
<td>Kitemoka</td>
<td>Chapakuran</td>
<td>Weinold (2012)</td>
</tr>
<tr>
<td>Bororo</td>
<td>Bororoan</td>
<td>Ochoa (1995)</td>
</tr>
<tr>
<td>Kadiwéu</td>
<td>Guaycuruan</td>
<td>Griffiths (2002)</td>
</tr>
<tr>
<td>Movima</td>
<td>Isolate</td>
<td>Judy and de Judy (1962)</td>
</tr>
<tr>
<td>Saraveka</td>
<td>Arawakan</td>
<td>de Créqui-Montfort and Rivet (1913)</td>
</tr>
<tr>
<td>Yurakaré</td>
<td>Isolate</td>
<td>Ribera (1991)</td>
</tr>
</tbody>
</table>
For Guaranian and Guarayan, Only 19 elements:

- 13 are also innovation in others languages in the family
- 8 are made up with TG material

For Siriono:

<table>
<thead>
<tr>
<th>Meaning</th>
<th>Sirionoan</th>
<th>Baure</th>
<th>Bésiro</th>
<th>Ayoreo</th>
<th>Saraveka</th>
<th>Yurakaré</th>
<th>Kitemoka</th>
<th>Movima</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mother</td>
<td>taïʔ</td>
<td>nen</td>
<td>piapá</td>
<td>daté, ité</td>
<td>mama-zudzani</td>
<td>meme</td>
<td>mama</td>
<td>ma’a</td>
</tr>
<tr>
<td>Sun</td>
<td>tenda</td>
<td>ses</td>
<td>sú</td>
<td>wedé</td>
<td>tayetara</td>
<td>toshta</td>
<td>putyo</td>
<td>tinno</td>
</tr>
<tr>
<td>Deer</td>
<td>k’ikʷá, jikan</td>
<td>kajaw; nor</td>
<td>ibó</td>
<td>erámoro</td>
<td>dyuluta</td>
<td>chirop</td>
<td>daujes</td>
<td></td>
</tr>
<tr>
<td>Back (n.)</td>
<td>kite</td>
<td>-chipi</td>
<td>chakú</td>
<td>guide-(boi)</td>
<td>dyuluta</td>
<td>totoyache</td>
<td>yaisaka</td>
<td></td>
</tr>
<tr>
<td>Dirt</td>
<td>e-túbi</td>
<td>topopkon</td>
<td>kufifu</td>
<td>topié</td>
<td></td>
<td></td>
<td>duducan</td>
<td></td>
</tr>
</tbody>
</table>

A needle in the haystack
Discussion of the results

No obvious sources for any loanword or areal word. Possible explanations for these outcomes:

• Contact source of innovation in a language outside of the area we examined
• Contact source of innovation now extinct
• Semantic shift or lexical replacement in source language made it difficult to find relevant forms
• Sirionoan lexical innovations are in fact not due to contact but internal changes

Our method seems useful in principle, even if it came up dry in this case.
Next steps

• Search in a larger sample of languages
• Integrate new data from the field:
  • On Guarasu, Guarayo and Siriono by Swintha Danielsen and Noé Gasparini
  • Eva-Maria Rößler will update Aché
  • Windy Daviet (MA student in Lyon) is starting a description of Bolivian Guaraní
• Phonological reconstruction
  + Phase 2: structural comparison (already started in the past year), collaborative work between Berkeley and Lyon
Acknowledgements

• **Diamantis Sellis** for crucial computational assistance:
  • automated binary coding
  • cognate set completeness and consistency checking scripts
• The following colleagues for generously sharing primary data:
  • **Sebastian Drude** (Awetî)
  • **Sérgio Meira** (Mawé, Tembé)
  • **Françoise Rose** (Emerillon)
  • **Eva-Maria Rößler** (Aché)
  • **Rosa Vallejos** (Kokama-Kokamilla)
• **Tammy Stark** for GIS assistance
• And the following Berkeley TG group alumni:
  • **Mike Roberts**
  • **Vivian Wauters**
• NSF DEL Award #0966499
• UC Berkeley Social Science Matrix 2013-2014 Grant
• France-Berkeley Fund
• LABEX ASLAN (ANR-10-LABX-0081) of Université de Lyon
References


Judy, Roberto and Judit Emerich de Judy. 1962. *Movima y Castellano. (Vocabularios Bolivianos, 1.).* Cochabamba, Bolivia: SIL.

Michael, Lev; Natalia Chousou-Polydouri; Keith Bartolomei; Erin Donnelly; Vivian Wauters; Sérgio Meira; and Zachary O’Hagan. to appear. A bayesian phylogenetic classification of tupí-guaraní. *LLAMES*. 


Tupí Guaraní languages - All subgroups
Lexicostatistics and Phylogenetics

• Lexicostatistical Methods (e.g. NeighborNet, SplitsTree)
  • Lexicostatistical methods do not evaluate evolutionary trees
  • Instead they compute a single number – % 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 less likely to be “fooled” by borrowing than lexicostatistical methods