A Phylogenetic Approach to the Study of the Tupí-Guaraní Language Family

Natalia Chousou-Polydouri, Vivian Wauters, Zachary O’Hagan, Keith Bartolomei, and Lev Michael

*University of California, Berkeley*

∼

Workshop on American Indigenous Languages
University of California, Santa Barbara
April 26, 2013
Introduction

Present a recent internal classification of Tupí-Guaraní (TG) languages based on the application of computational phylogenetic methods to a TG lexical dataset.

- Overview of TG languages and internal classification
- TG lexical dataset
- Phylogenetic methods
- Results
- Comparison with previous classifications

Current project grows out of efforts to understand the relationship between Omagua (& Kokama) and the other TG languages.
Tupí classification

- Universal agreement among specialists that TG forms a subgroup within the Tupí stock (Kaufman 1994; Campbell 1997; Jensen 1999; Rodrigues 1986, 1999).

- Also now (cf. Rodrigues (1958)) universal agreement that Aweti and Sateré-Mawé are – in that order – the Tupian languages most closely related to TG (Rodrigues and Dietrich 1997; Drude 2006, 2011; Corrêa da Silva 2007, 2010; Kamaiurá 2012).

Figure 1: Tupí Classification (Drude 2011)
Tupí-Guaraní classification

- The most influential classification groups the ~40 living and extinct TG languages into eight sub-groups (Rodrigues 1985).
  - The empirical basis of this classification is unclear (more below).
- Limited efforts to group the eight sub-groups into larger higher order groups (Lemle 1971; Mello 2000, 2002; Rodrigues and Cabral 2002).
  - results are coarse-grained and contradictory
  - methods and data employed are unclear
Rodrigues (1985) Classification of Tupí-Guaraní languages

<table>
<thead>
<tr>
<th>PTG</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
<th>VII</th>
<th>VIII</th>
</tr>
</thead>
<tbody>
<tr>
<td>*C#</td>
<td>*C# &gt; Ø</td>
<td>*C# &gt; Ø</td>
<td>*C# &gt; Ø</td>
<td>*C# &gt; Ø</td>
<td>*C# &gt; Ø</td>
<td>*C# &gt; Ø</td>
<td>*C# &gt; Ø</td>
<td>*C# &gt; Ø</td>
</tr>
<tr>
<td>*tf</td>
<td>*tf &gt; ts ~ s</td>
<td>*tf &gt; ts ~ s</td>
<td>*tf &gt; ts ~ s</td>
<td>*tf &gt; ts ~ s</td>
<td>*tf &gt; ts ~ s</td>
<td>*tf &gt; ts ~ s</td>
<td>*tf &gt; ts ~ s</td>
<td>*tf &gt; ts ~ s</td>
</tr>
<tr>
<td>*ts</td>
<td>*ts &gt; h ~ Ø</td>
<td>*ts &gt; h ~ Ø</td>
<td>*ts &gt; h ~ Ø</td>
<td>*ts &gt; h ~ Ø</td>
<td>*ts &gt; h ~ Ø</td>
<td>*ts &gt; h ~ Ø</td>
<td>*ts &gt; h ~ Ø</td>
<td>*ts &gt; h ~ Ø</td>
</tr>
<tr>
<td>*pw</td>
<td>*pw &gt; kw ~ k</td>
<td>*pw &gt; kw ~ k</td>
<td>*pw &gt; kw ~ k</td>
<td>*pw &gt; kw ~ k</td>
<td>*pw &gt; kw ~ k</td>
<td>*pw &gt; kw ~ k</td>
<td>*pw &gt; kw ~ k</td>
<td>*pw &gt; kw ~ k</td>
</tr>
<tr>
<td>*pj</td>
<td>*pj &gt; tf ~ f</td>
<td>*pj &gt; tf ~ f</td>
<td>*pj &gt; tf ~ f</td>
<td>*pj &gt; tf ~ f</td>
<td>*pj &gt; tf ~ f</td>
<td>*pj &gt; tf ~ f</td>
<td>*pj &gt; tf ~ f</td>
<td>*pj &gt; tf ~ f</td>
</tr>
<tr>
<td>*j</td>
<td>*j &gt; tf ~ ts</td>
<td>*j &gt; tf ~ ts</td>
<td>*j &gt; tf ~ ts</td>
<td>*j &gt; tf ~ ts</td>
<td>*j &gt; tf ~ ts</td>
<td>*j &gt; tf ~ ts</td>
<td>*j &gt; tf ~ ts</td>
<td>*j &gt; tf ~ ts</td>
</tr>
<tr>
<td>STRESS</td>
<td>*σ &gt; σ#</td>
<td>*σ &gt; σ#</td>
<td>*σ &gt; σ#</td>
<td>*σ &gt; σ#</td>
<td>*σ &gt; σ#</td>
<td>*σ &gt; σ#</td>
<td>*σ &gt; σ#</td>
<td>*σ &gt; σ#</td>
</tr>
</tbody>
</table>

**LANGS.**

<table>
<thead>
<tr>
<th>Old Guarani</th>
<th>Guarayú</th>
<th>Tupinambá</th>
<th>Tapirapé</th>
<th>Kayabí</th>
<th>Parintintin</th>
<th>Kamaiurá</th>
<th>Takunhapé</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mbyá</td>
<td>Sirionó</td>
<td>Tupí Austral</td>
<td>Avá-Canoeiro</td>
<td>Asurini do X.</td>
<td>Tupí-Kawahib</td>
<td>Apiaká</td>
<td>Wayampí</td>
</tr>
<tr>
<td>Xetá</td>
<td>Jorá</td>
<td>Nheengatú</td>
<td>Asurini do T.</td>
<td>Surúí</td>
<td>Guajajara</td>
<td>Guajá</td>
<td>Wayampipukú</td>
</tr>
<tr>
<td>Ñandeua</td>
<td></td>
<td>Kokama</td>
<td>Parakaná</td>
<td></td>
<td></td>
<td></td>
<td>Eméllion</td>
</tr>
<tr>
<td>Kaowá</td>
<td></td>
<td>Kokamilla</td>
<td>Guajajára</td>
<td></td>
<td></td>
<td></td>
<td>Anamaye</td>
</tr>
<tr>
<td>P. Guarani</td>
<td>Guayakí</td>
<td>Omagua</td>
<td>Tembé</td>
<td></td>
<td></td>
<td></td>
<td>Anambé</td>
</tr>
<tr>
<td></td>
<td>(Aché)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Turiwára</td>
</tr>
<tr>
<td>Tapieté</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Guajá</td>
</tr>
<tr>
<td>Chiriguano (Avá)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Urubú-Ka’apor</td>
</tr>
<tr>
<td>Izoceño (Chané)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 2:** Tupí-Guaraní Subgroups and Sound Changes (Rodrigues 1985)
Rodrigues (1985) classification of Tupí-Guaraní languages
Data Sources and Harvesting

- The TG Comparative Lexical Database (TGCLD) includes data from 30 TG languages and 2 non-TG Tupian languages (Sateré-Mawé and Awetí), which serve as outgroup languages.
- These 30 languages constitute nearly all TG languages for which significant lexical data exists.
- Data was harvested by Keith Bartolomei, Natalia Chousou-Polydouri, Lev Michael, Zachary O’Hagan, Mike Roberts, and Vivian Wauters from:
  - dictionaries
  - phonological descriptions
  - grammatical descriptions
- This work is based on a 539-item comparative list of cross-linguistically basic and areally-appropriate vocabulary
  - e.g., ‘jaguar’, ‘manioc’, ‘peccary’
### Lexical Coverage

<table>
<thead>
<tr>
<th>Language</th>
<th>Coverage</th>
<th>Comparison to Other Language</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aché</td>
<td>76%</td>
<td>Kokama</td>
<td>90%</td>
</tr>
<tr>
<td>Anambé</td>
<td>32%</td>
<td>Mbyá</td>
<td>83%</td>
</tr>
<tr>
<td>Apiaká</td>
<td>18%</td>
<td>Ñandéva</td>
<td>22%</td>
</tr>
<tr>
<td>Araweté</td>
<td>55%</td>
<td>Omagua</td>
<td>91%</td>
</tr>
<tr>
<td>Asuriní do Tocantins</td>
<td>74%</td>
<td>Parakanã</td>
<td>78%</td>
</tr>
<tr>
<td>Asuriní do Xingú</td>
<td>53%</td>
<td>Paraguayan Guarani</td>
<td>94%</td>
</tr>
<tr>
<td>Avá-Canoeiro</td>
<td>55%</td>
<td>Pauserna</td>
<td>60%</td>
</tr>
<tr>
<td>Awetí (non-TG)</td>
<td>76%</td>
<td>Sateré-Mawé (non-TG)</td>
<td>43%</td>
</tr>
<tr>
<td>Chiriguano</td>
<td>82%</td>
<td>Sirionó</td>
<td>83%</td>
</tr>
<tr>
<td>Emérillon</td>
<td>79%</td>
<td>Tapieté</td>
<td>86%</td>
</tr>
<tr>
<td>Guajá</td>
<td>47%</td>
<td>Tapirapé</td>
<td>72%</td>
</tr>
<tr>
<td>Guarayú</td>
<td>81%</td>
<td>Tembé</td>
<td>92%</td>
</tr>
<tr>
<td>Ka’ápor</td>
<td>75%</td>
<td>Tupinambá</td>
<td>93%</td>
</tr>
<tr>
<td>Kaiowá</td>
<td>41%</td>
<td>Wayampí</td>
<td>88%</td>
</tr>
<tr>
<td>Kamaiurá</td>
<td>59%</td>
<td>Xetá</td>
<td>34%</td>
</tr>
<tr>
<td>Kayabí</td>
<td>62%</td>
<td>Yuki</td>
<td>82%</td>
</tr>
</tbody>
</table>
Building and Coding Cognate Sets

- 1,324 non-trivial cognate sets (4,152 total sets)
- Constructed and reviewed collaboratively:
  - making use of impressionistic sound correspondences
  - items which have undergone semantic shift were included in the cognate sets to which they belong
    - Ex.: Omagua *miara* ‘monkey’ moved to HUNT etymon
- Treatment of compounds:
  - We coded compounding strategies as a character (e.g. ‘star’ = ‘moon’ + ‘fire’).
  - Words present as part of a compound were also coded as present in the relevant cognate sets.
## Building and Coding Cognate Sets

### Binary Coding

<table>
<thead>
<tr>
<th>Cognate Sets/Language</th>
<th>Guarani Paraguai</th>
<th>Kalowá</th>
<th>Mbyá</th>
<th>Ñandéva</th>
<th>Chiriguano</th>
<th>Sirionó</th>
<th>Yuki</th>
<th>Kokama</th>
<th>Tupinambá</th>
<th>Tsimiré</th>
<th>Parakaná</th>
<th>Satere Mawe</th>
</tr>
</thead>
<tbody>
<tr>
<td>head</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HEAD1</td>
<td>&lt;akã&gt;</td>
<td>&lt;akã&gt;</td>
<td>&lt;akã&gt;</td>
<td>[ãkã] [2]</td>
<td>/ãkã/</td>
<td>BRAIN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HEAD2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HEAD3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HEAD4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRAIN1</td>
<td>&lt;apitu'ù&gt;</td>
<td>&lt;apitu'ù&gt;</td>
<td>/apitu/</td>
<td>&lt;eitoš&gt;</td>
<td>&lt;itö&gt; (head)[2]</td>
<td>HEAD</td>
<td>&lt;apituma&gt;</td>
<td>(miolo, polpo)</td>
<td>&lt;apita'am&gt;</td>
<td>[tamîkim]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRAIN2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table

<table>
<thead>
<tr>
<th>Cognate Sets/Language</th>
<th>Guarani Paraguai</th>
<th>Kalowá</th>
<th>Mbyá</th>
<th>Ñandéva</th>
<th>Chiriguano</th>
<th>Sirionó</th>
<th>Yuki</th>
<th>Kokama</th>
<th>Tupinambá</th>
<th>Tsimiré</th>
<th>Parakaná</th>
<th>Satere Mawe</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEAD1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>HEAD2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HEAD3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HEAD4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BRAIN1</td>
<td>1</td>
<td>?</td>
<td>1</td>
<td>?</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>?</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>BRAIN2</td>
<td>0</td>
<td>?</td>
<td>0</td>
<td>?</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>?</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Phylogenetic Methods (MrBayes 3.2)

- We used a restriction site model
  - Gamma distributed rates
  - Starts with uniform prior for cognate gain/loss, with allowance for difference in rates to emerge
  - Also allows for difference in loss/gain per cognate (simulates difference across etyma for rate of change)

- Analysis conducted with four independent runs
  - 10 million generations each, sampled every 1,000 generations
  - the chains spend time in (and therefore sample) the posterior distribution of trees proportionally to their posterior probability

- Stationarity
  - adequate sampling of the posterior distribution
  - convergence (independent runs sampling the same posterior distribution) assessed with Tracer v1.5 and MrBayes 3.2
Results

Phylogenetic Results (MrBayes 3.2)

- All runs converged with an average standard deviation of split frequencies of 0.003
- Analysis produced single majority-rule consensus tree, with posterior probabilities
- Initial 25% of trees sampled during analysis discarded as ‘burn-in’
- Asymmetry of loss vs. gain of cognates was 16.95:1
  - indicating ‘Dollo-like’ model
  - cognates emerge once in the tree but can be lost independently in different languages
  - low level of borrowing
A Phylogenetic Approach to the Study of the Tupí-Guaraní Language Family

Results

Bayesian TG Classification
Conservative TG Classification

A Phylogenetic Approach to the Study of the Tupí-Guaraní Language Family

Classification

Results
A Phylogenetic Approach to the Study of the Tupí-Guaraní Language Family

Results

Classification

Northern Split

Tupí-Guaraní

Central

Diasporic

Southern Guaranian

Wayampí
Emérillon
Avá-Canoeiro
Kaapor
Guajá
Kayabi
Apiaká
Aché
Kameirurá
Tapirapé
Tupinambá
Guarayú
Pauana
Sirionó
Yuki
Tepeté
Paraguayan Guarani
Mbyá
Kalowá
Nandéva
Xetá
Geographic Distribution of Northern Split
“Second” Split

A Phylogenetic Approach to the Study of the Tupí-Guaraní Language Family

Results

Classification
Geographic Distribution “Second” Split
Central Split

A Phylogenetic Approach to the Study of the Tupí-Guaraní Language Family

Results

Classification

Central Split
Geographic Distribution of Central Split
Results

Classification

Tembé Split

[Diagram showing the classification of language families and groups, with branches for Central, Diasporic, Southern, and Guaranian, including labels for various language groups such as Wayampi, Kayabi, Arawete, Tembé, Paraguayan Guarani, and Xetá.]
Geographic Distribution Tembé Split
Diasporic Split

A Phylogenetic Approach to the Study of the Tupí-Guaraní Language Family

Results

Classification

Wayampí  Emérillon  Kayabí  Kaíñor  Guajá  Araweté  Tembé

Central

Tupí-Guaraní

Diasporic

Guarayú  Pausena  Siriono  Yuki  Tembé  Tupinambá  Omagua  Kokama  Tapete  Chiriguano

Southern

Guaranian

Tupinambá  Omagua  Kokama  Tapete  Chiriguano  Paraguay Guaraní  Mbyá  Kaiwá  Ñandeva  Xetá
Geographic Distribution of Diasporic Split
Southern Split

A Phylogenetic Approach to the Study of the Tupí-Guaraní Language Family

- Results
- Classification

Southern Split

- Wayampí
- Emérillon
- Avá-Canoeiro
- Kaapor
- Guaëa
- Kayabí
- Apiaká
- Aché
- Kamaiurá
- Tapirapé
- Perakarâ
- Asurini do Tocantins
- Asurini do Xingu
- Tupanambá
- Araweté
- Tembé
- Guaraní
- Guaraní
- Guaraní
- Guaraní
- Guaraní

Diasporic

- Tupí-Guaraní
- Tupí-Guaraní
- Tupí-Guaraní
- Tupí-Guaraní
- Tupí-Guaraní

Central

- Guaraí
- Pauana
- Siriona
- Yuki
- Tupanambá
- Omega
- Kokama
- Tapiete
- Paraguayan Guarani
- Mbya
- Kiatwa
- Nande
- Xetá
Geographic Distribution Southern Split
Comparison with Rodrigues and Cabral (2002)
Comparison with Rodrigues and Cabral (2002)

- Group I minus Aché = Southern
- Group II+III = paraphyletic group of Diasporic minus Southern
  - Would group together based on shared retentions
- Group IV+V+VI+VII ≈ Central
  - R&C include Avá-Canoeiro & Tembé in Group IV, but here they form a group with Ka’apor & Guajá and an independent first-order split, respectively
  - For Group IV, V & VI languages we have data on, their grouping mirrors R&C exactly
  - For R&C Kamaiurá is a paraphyletic group to their IV, V & VI, whereas here it forms independent branch of Central
- Group VIII = two first-order splits plus Avá-Canoeiro
- No motivation for excluding Pauserna (Rodrigues 1958, 1985)
Comparison with Previous Classifications

Conservative TG Classification
Conclusion and Future Directions

Summary
- We have presented an internal classification of TG based on a 539-item comparative list that yielded 1,324 non-trivial cognate sets.
  - broadly consistent with previous internal classifications of TG

Future Directions
- Add more lexical data:
  - to included languages (e.g., Kayabí, Kamaiurá)
  - from unincluded languages (e.g., Juma, Tupí-Kawahib cluster, Zo’é)
- Multi-state semantic coding
- Morphological characters
- Phonological characters
- Further phylogenetic analyses:
  - BEAST
  - Estimation of divergence times
- Comparative Method
Acknowledgments

- Diamantis Sellis
  - automated binary coding
  - scripts to check completeness and consistency of cognate sets
- Sebastian Drude, Françoise Rose and Rosa Vallejos Yopán
  - Unpublished lexical data on Awetí, Emérillon and Kokama-Kokamilla
- Noé Gasparini
  - Access to Yuki and Anambé data (Garland 1978; Silva Julião 2005)
- Mike Roberts
  - Paraguayan Guarani data harvesting
- UC Berkeley *dhworom* historical linguistics working group
- National Science Foundation DEL award #0966499: Collaborative Research: Kokama-Kokamilla (cod) and Omagua (omg): Documentation, Description, and (Non-)Genetic Relationships
References I


References III


Lemle (1971)

Figure 1: Tupí-Guaraní Subgrouping (Lemle 1971)
Walker et al. (2012)
A Phylogenetic Approach to the Study of the Tupí-Guaraní Language Family

Conclusion

Geographic Spread
Comparison with archeology and physical anthropology

- Geographically coherent model of language spread based on internal classification of family
- Model of spread consistent with archeological work (Noelli 1998, 2008):
  - placing Tupí homeland between Madeira and Tapajós Rivers;
  - positing a coastal expansion for the southern spread of TG.
- Also consistent with craniometric work in physical anthropology linking (southern) Guaranian peoples with populations near the mouth of the Amazon (Neves et al. 2011).