A Bayesian phylogenetic internal classification of the Tupí-Guaraní family

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

*dhworom
February 14, 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
  - Evaluating the internal classification
    - Linguistic diversification vs. geography, archeology, and physical anthropology
    - Comparison with previous classifications
- Current project grows out of efforts to understand the relationship between Omagua (& Kokama) and the other TG languages.
- Next step: application of Comparative Method.
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 Awetí 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 $\sim 40$ 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
A Bayesian phylogenetic internal classification of the Tupí-Guaraní family

Introduction

Rodrigues (1984) classification of Tupí-Guaraní languages

Table 1: Tupí-Guaraní Subgroups and Sound Changes (Rodrigues 1984/1985)

<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; h</td>
<td>*tf &gt; h~ Ø</td>
<td>*tf &gt; h</td>
<td>*tf &gt; h~ Ø</td>
<td>*tf &gt; h~ Ø</td>
<td>*tf &gt; h~ Ø</td>
</tr>
<tr>
<td>*ts</td>
<td>*ts &gt; ts ~ s</td>
<td>*ts &gt; ts~ s</td>
<td>*ts &gt; ts ~ s</td>
<td>*ts &gt; ts ~ s</td>
<td>*ts &gt; ts ~ s</td>
<td>*ts &gt; ts ~ s</td>
<td>*ts &gt; ts ~ s</td>
<td>*ts &gt; ts ~ s</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; ts ~ f</td>
<td>*pj &gt; ts ~ f</td>
<td>*pj &gt; ts ~ f</td>
<td>*pj &gt; ts ~ f</td>
<td>*pj &gt; ts ~ f</td>
<td>*pj &gt; ts ~ f</td>
<td>*pj &gt; ts ~ f</td>
<td>*pj &gt; ts ~ f</td>
</tr>
<tr>
<td>*j</td>
<td>*j &gt; ts ~ s</td>
<td>*j &gt; ts ~ s</td>
<td>*j &gt; ts ~ s</td>
<td>*j &gt; ts ~ s</td>
<td>*j &gt; ts ~ s</td>
<td>*j &gt; ts ~ s</td>
<td>*j &gt; ts ~ s</td>
<td>*j &gt; ts ~ s</td>
</tr>
<tr>
<td>STRESS</td>
<td>Ø</td>
<td>Ø</td>
<td>Ø</td>
<td>Ø</td>
<td>Ø</td>
<td>Ø</td>
<td>Ø</td>
<td>Ø</td>
</tr>
</tbody>
</table>

Table 1: Tupí-Guaraní Subgroups and Sound Changes (Rodrigues 1984/1985)

<table>
<thead>
<tr>
<th>LANGS.</th>
<th>Old Guarani</th>
<th>Guaraní</th>
<th>Tupí-Austral</th>
<th>Tapirapé</th>
<th>Kayabi</th>
<th>Parintintin</th>
<th>Kamaiurá</th>
<th>Takunhapé</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mbyá</td>
<td>Sirionó</td>
<td>Tupinambá</td>
<td>Avá-Canoeiro</td>
<td>Asurini do X.</td>
<td>Tupil-Kawahib</td>
<td>Apiaká</td>
<td>Kamaiurá</td>
<td>Takunhapé</td>
</tr>
<tr>
<td>Ñandeva</td>
<td>Jorá</td>
<td>Nheengatú</td>
<td>Asurini do T.</td>
<td>Suruí</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kaiová</td>
<td></td>
<td>Kokama</td>
<td>Parakaná</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P. Guarani</td>
<td>Guayakí (Aché)</td>
<td>Kokamilla</td>
<td>Guajajára</td>
<td>Tembé</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tapite</td>
<td></td>
<td>Omagua</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chiriguanó (Avá)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Izoceño (Chané)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Rodrigues (1984) classification of Tupí-Guaraní languages
Data Sources and Harvesting

- The TG Comparative Lexical Database (TGCLD) includes data from 28 TG languages and 2 non-TG Tupian languages (Sateré-Mawé and Awetí), which serve as outgroup languages.
- These 28 languages constitute all the TG languages for which significant lexical data exists.
- Data was harvested by Keith Bartolomei, Natalia Chousou-Polydouri, Lev Michael, Zach O’Hagan, Mike Roberts, and Vivian Wauters (who additionally managed the TGCLD) from:
  - dictionaries
  - phonological descriptions
  - grammatical descriptions
- We also harvested morphological data, but we do not analyze that data here.
This work is based on a 572-item comparative list of cross-linguistically basic and areally-appropriate vocabulary (e.g. ‘jaguar’, ‘manioc’).
### Data & Methodology

<table>
<thead>
<tr>
<th>Language</th>
<th>Coverage</th>
<th>Comparison</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apiaká</td>
<td>30%</td>
<td>Mbyá</td>
<td>84%</td>
</tr>
<tr>
<td>Araweté</td>
<td>55%</td>
<td>Ñandeva</td>
<td>21%</td>
</tr>
<tr>
<td>Assuriní do Tocantins</td>
<td>77%</td>
<td>Omagua</td>
<td>91%</td>
</tr>
<tr>
<td>Assuriní do Xingú</td>
<td>57%</td>
<td>Parakanã</td>
<td>78%</td>
</tr>
<tr>
<td>Avá-Canoeiro</td>
<td>55%</td>
<td>Paraguayan Guaraní</td>
<td>95%</td>
</tr>
<tr>
<td>Awetí (non-TG)</td>
<td>44%</td>
<td>Parintintín</td>
<td>80%</td>
</tr>
<tr>
<td>Chiriguano</td>
<td>81%</td>
<td>Sateré-Mawé (non-TG)</td>
<td>44%</td>
</tr>
<tr>
<td>Emerillón</td>
<td>46%</td>
<td>Sirionó</td>
<td>83%</td>
</tr>
<tr>
<td>Guajá</td>
<td>47%</td>
<td>Tapieté</td>
<td>87%</td>
</tr>
<tr>
<td>Guarayú</td>
<td>82%</td>
<td>Tapirapé</td>
<td>73%</td>
</tr>
<tr>
<td>Ka’apor</td>
<td>77%</td>
<td>Tembé</td>
<td>93%</td>
</tr>
<tr>
<td>Kaiowá</td>
<td>40%</td>
<td>Tupinambá</td>
<td>94%</td>
</tr>
<tr>
<td>Kamaiurá</td>
<td>59%</td>
<td>Wayampí</td>
<td>91%</td>
</tr>
<tr>
<td>Kayabí</td>
<td>63%</td>
<td>Xetá</td>
<td>34%</td>
</tr>
<tr>
<td>Kokama</td>
<td>90%</td>
<td>Yuki</td>
<td>56%</td>
</tr>
</tbody>
</table>

**Table 1: Coverage by language of TG comparative list**
Building Cognate Sets

- We constructed 1400 non-singleton cognates sets from the comparative list data (2764 total sets).
- Initially, we formed cognate sets out of individually-assigned sections of the comparative list.
- We discussed sets in weekly meetings, often making use of sound correspondences we noticed in the dataset.
- Later we each reviewed the entire set in detail (semantic shift being a major issue), discussing doubts about particular cognate sets in weekly meetings.
Character Coding

- The results we present today are derived from binary coding.
- Items which have undergone semantic shift were included in the cognate sets to which they belong.
  - Ex.: Omagua *miara* ‘monkey’ moved to **MEAT** 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.
Phylogenetic Methods

- Bayesian phylogenetic model (MrBayes)
  - 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 MCMC (Metropolis-Coupled Markov Chain Monte Carlo) runs. Each run consists of a cold - sampling chain and 3 hot - exploring chains.
    - 10 million generations each, sampled every 5000 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) and convergence (independent runs sampling the same posterior distribution) assessed with Tracer v1.5 and MrBayes 3.2
Phylogenetic Results

- Bayesian phylogenetic model (MrBayes)
  - all runs converged
  - Majority rule consensus tree, with posterior probabilities
  - 25% discarded as burn-in
  - Asymmetry of loss vs. gain of cognates was 13.7:1
Bayesian TG Classification

![Bayesian phylogenetic internal classification of the Tupí-Guaraní family](image-url)

Results
A Bayesian phylogenetic internal classification of the Tupí-Guaraní family

Results

Conservative TG Classification

![Tupí-Guaraní Subgrouping](image)

Figure 1: Tupí-Guaraní Subgrouping (MrBayes)
A Bayesian phylogenetic internal classification of the Tupí-Guaraní family

Results

1st and 2nd order TG subgroups

Figure 1: Tupí-Guaraní Higher Level Subgrouping (MrBayes)
A Bayesian phylogenetic internal classification of the Tupí-Guaraní family

Results

Geographical distribution of 1st and 2nd order subgroups
A Bayesian phylogenetic internal classification of the Tupí-Guaraní family

Results

Central subgroup

Figure 1: Tupí-Guaraní Subgrouping – ‘Central’ (MrBayes)
Geographical distribution of Central subgroups
Peripheral subgroup

Figure 1: Tupí-Guaraní Subgrouping – ‘Peripheral’ (MrBayes)
Geographical distribution of Peripheral subgroups
A Bayesian phylogenetic internal classification of the Tupí-Guaraní family

Results

16th century Omagua-Kokama & Tupinambá distribution
TG internal classification and geographical dispersion

- Phylogenetic results present challenges for evaluation: there is little argumentation to examine.
- In this section we use non-linguistic evidence to evaluate the plausibility of our internal classification:
  - Use TG internal classification to infer a plausible language spread scenario
  - Compare with non-linguistic evidence to assess plausibility of spread model, and hence internal classification
A Bayesian phylogenetic internal classification of the Tupí-Guaraní family

PTG

Discussion

TG internal classification and geographical dispersion
A Bayesian phylogenetic internal classification of the Tupí-Guaraní family

---

**Discussion**

TG internal classification and geographical dispersion

1st order split

---
A Bayesian phylogenetic internal classification of the Tupí-Guaraní family

Discussion

TG internal classification and geographical dispersion

Main Group split
A Bayesian phylogenetic internal classification of the Tupí-Guaraní family

Discussion

TG internal classification and geographical dispersion

Peripheral Group split
A Bayesian phylogenetic internal classification of the Tupí-Guaraní family

Discussion

TG internal classification and geographical dispersion

Diasporic Group split

![Map of Tupí-Guaraní family classification and geographical dispersion]
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).
Comparison with previous classifications

- Another way to evaluate the plausibility of our internal classification of TG is to compare it with previous influential classifications.
  - Presumably TG specialists steeped in the data were more often right than wrong about lower- and mid-level groupings.
  - Our classification is thus more credible to the degree that our classification reproduces the major features of previous classifications.
A Bayesian phylogenetic internal classification of the Tupí-Guaraní family

Discussion

Comparison with previous classifications

Lemle (1971)

Figure 1: Tupí-Guaraní Subgrouping (Lemle 1971)
Comparison with Lemle (1971)

- Constructs 212 cognate sets for ten languages, but does ‘word-by-word’ reconstruction.
- Lemle’s Group I+II = Southern
- Lemle’s Group I+II+III = Expansionist (∼ Peripheral, since no discussion of Tembé or Northern Peripheral langs)
- Lemle’s right branch does not correspond tidily to any aspect of our classification
A Bayesian phylogenetic internal classification of the Tupí-Guaraní family

- Discussion
- Comparison with previous classifications

Rodrigues and Cabral (2002)
Comparison with Rodrigues and Cabral (2002)

- **Group I** = Guaranian
- **Group II+III** = paraphyletic group in Peripheral, above Guaranian and below Northern
  - Would group together based on shared retentions
- **Group IV+V+VI+VII** ≈ Central
  - R&C include Avá-Canoeiro in Group IV, while we consider it a top-level sister of our Main group
  - R&C include Tembé in Group IV, while we consider it a top-level sister, within Peripheral, of the Southern and Expansionist groups
  - We reverse the positions of Kamaiurá (Group VII) and the East Central group (group VI)
- **Group VIII** = North Peripheral + Guajá-Ka’ápor (a top level sister to our Main group)
Conclusion and Future Directions

Summary

- We have presented an internal classification of TG based on a 572-item comparative list that yielded 1400 non-trivial cognate sets.
  - Broadly consistent with archeological and physical anthropological evidence.
  - Broadly consistent with previous internal classifications of TG.

Future Directions

- Multi-state semantic coding
- Morphological characters
- Phonological characters
- Further phylogenetic analyses:
  - BEAST
  - Estimation of divergence times
- Comparative Method (!)
Thanks to:

- Diamantis Sellis
  - assistance with automating binary character conversion of cognate sets
- Mike Roberts
  - Paraguayan Guaraní data harvesting
A Bayesian phylogenetic internal classification of the Tupí-Guaraní family

Conclusion

Walker et al. (2012)
References I


References II


References III


