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

$\sim$

Linguistic Society of America
Annual Meeting
Portland, Oregon
January 10, 2015

Introduction I

- 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

Introduction II

- 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

Introduction IV

- 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

G&A Set Construction and Coding I

- 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' $\rightarrow$ '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

	OMG	KK	TPN	TPR	PAR
woman	wajnu	wajna	kujã	koʃɨ̃	koʃo
sister	kunia	kuɲa	ɨker	ɨket	ɨker
G&A (Multistate)					
woman	1	1	2	2	2
sister	1	1	2	2	2
G&A (Binary Recoding)					
WOMAN1 (**wajnua*)	1	1	0	0	0
WOMAN2 (**kujã*)	**0**	**0**	**1**	**1**	**1**
SISTER1 (**kujã*)	**1**	**1**	**0**	**0**	**0**
SISTER2 (**ɨker*)	0	0	1	1	1
QC (Binary)					
WOMAN1 (**wajnua*)	1	1	0	0	0
WOMAN2 (**kujã*)	**1**	**1**	**1**	**1**	**1**
SISTER1 (**ɨker*)	0	0	1	1	1

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

- Tupí-Guaraní is a well-established subgroup of the larger Tupí stock (Campbell 1997; Jensen 1999; Kaufman 1994, 2007; Rodrigues 1986, 1999; Rodrigues and Cabral 2012)
- First phylogenetic exploration of Tupí: Galúcio et al. (2013)

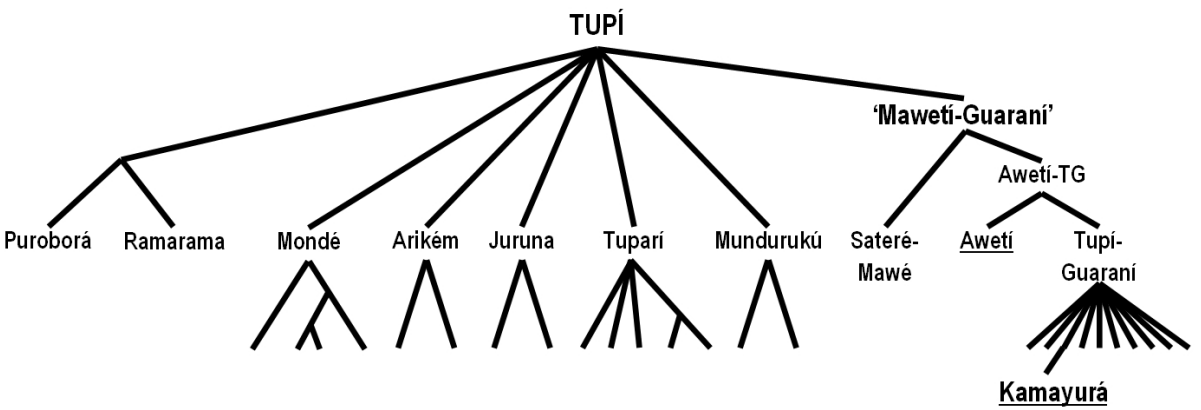


Figure 1: Tupí Classification (Drude 2011)

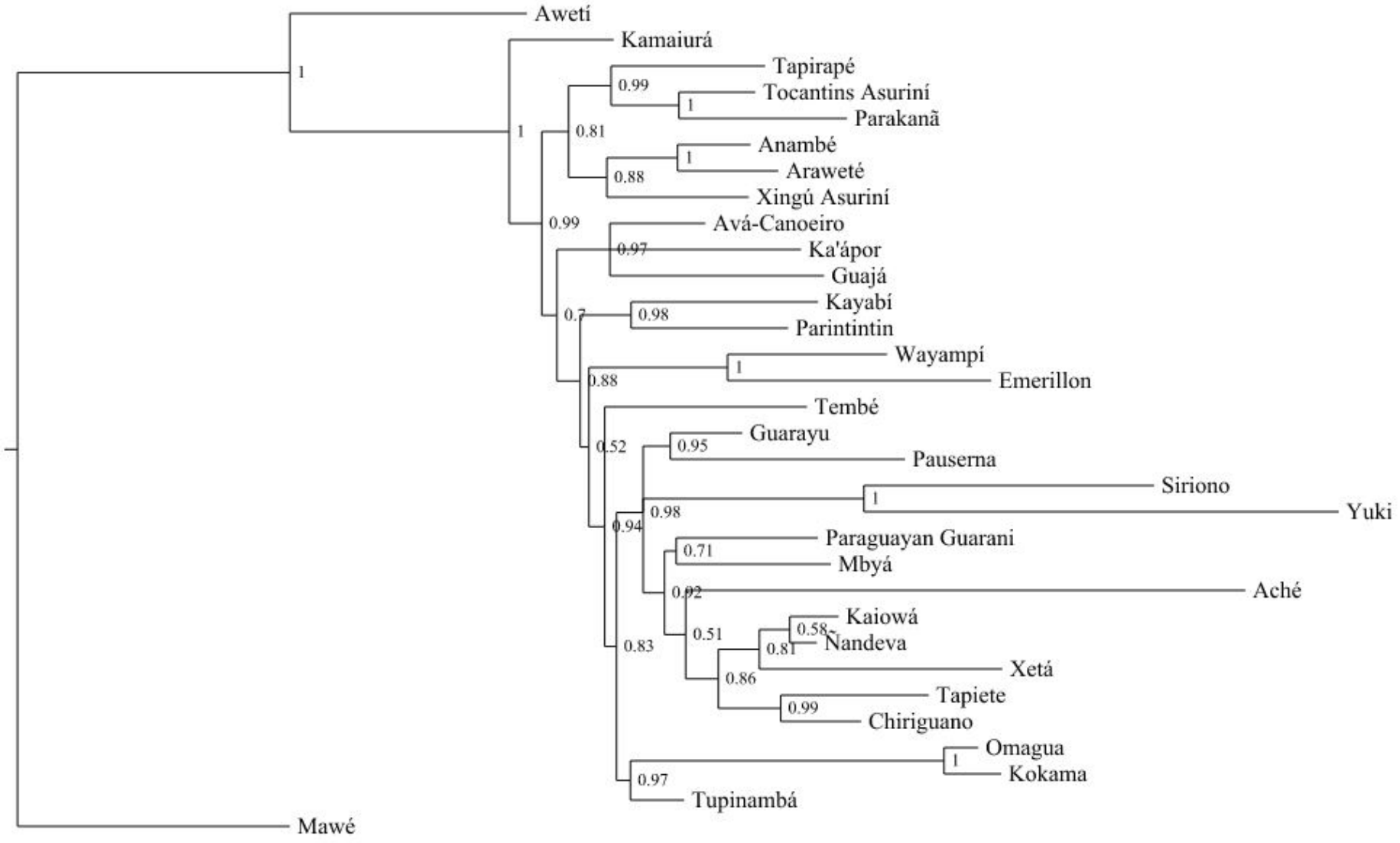
Data Harvesting

- There are ∼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%

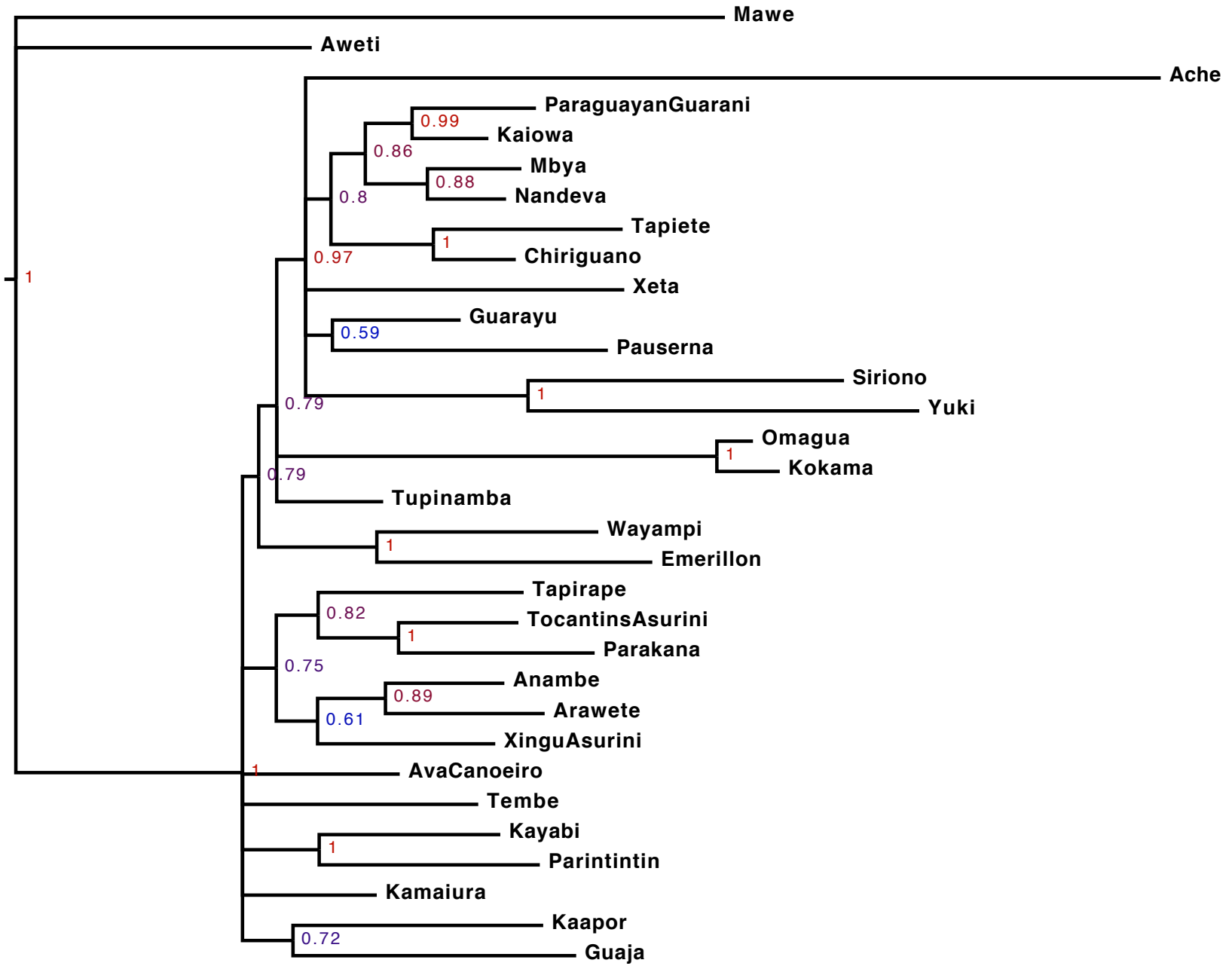
Lexical Coverage

Aché	85%	Ñandeva	20%
Anambé	31%	Omagua	89%
Araweté	55%	Parakanã	75%
Avá-Canoeiro	51%	Paraguayan Guarani	94%
Awetí	76%	Parintintin	85%
Chiriguano	80%	Pauserna	58%
Emerillon	77%	Siriono	82%
Guajá	45%	Tapiete	84%
Guarayu	86%	Tapirapé	69%
Ka'apor	83%	Tembé	98%
Kaiowá	39%	Tocantins Asuriní	83%
Kamaiurá	75%	Tupinambá	94%
Kayabí	59%	Wayampí	89%
Kokama	89%	Xetá	33%
Mawé	80%	Xingú Asuriní	50%
Mbyá	83%	Yuki	80%

Tupí-Guaraní Classification: QC Coding



Tupí-Guaraní Classification: G&A Coding



Comparison of Classifications I

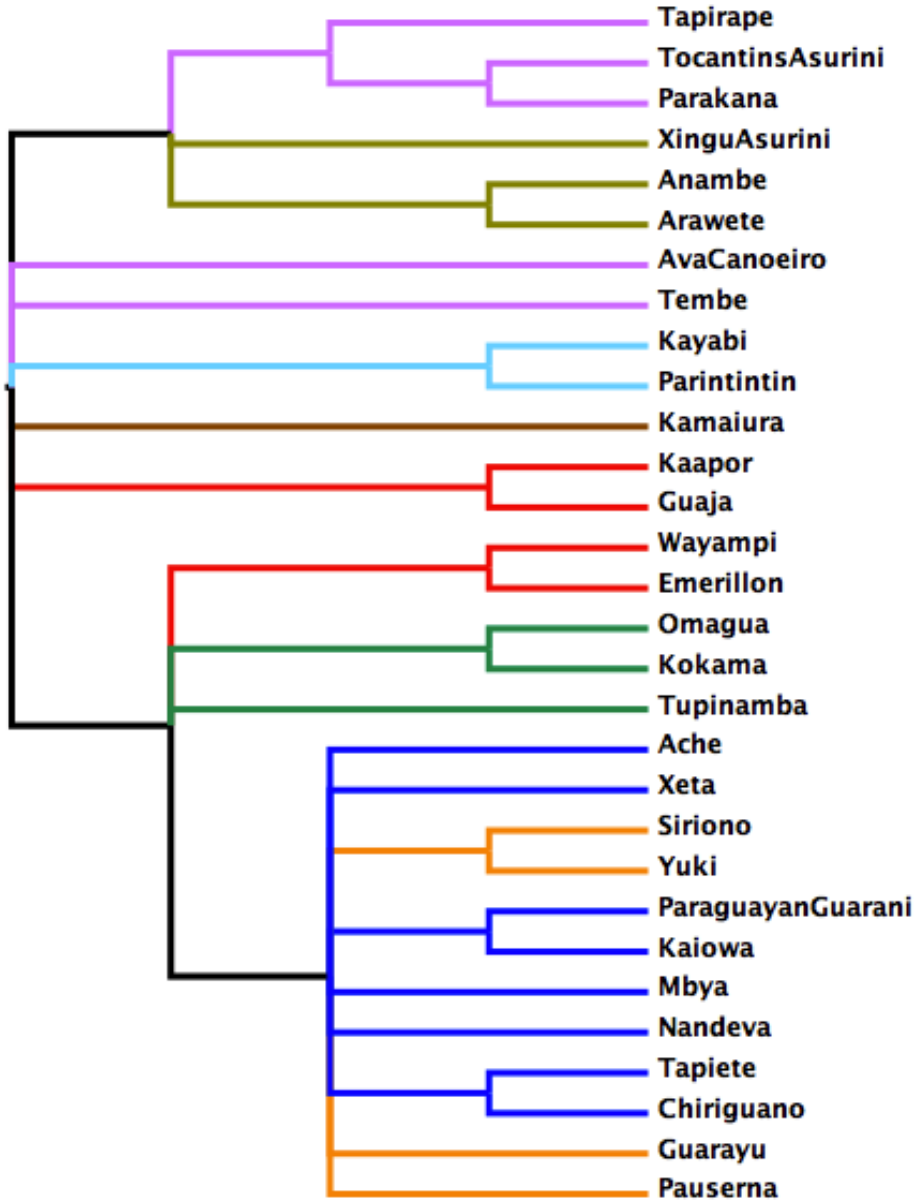
QC	G&A
More higher-level articulation	Less higher-level articulation
More unique nodes	Fewer unique nodes
Kamaiurá is sister to 'Nuclear-TG'	Kamaiurá is part of polytomy
Member of subgroup (Tupinambá, Avá-Canoeiro)	Merged with polytomy (Tupinambá, Avá-Canoeiro)

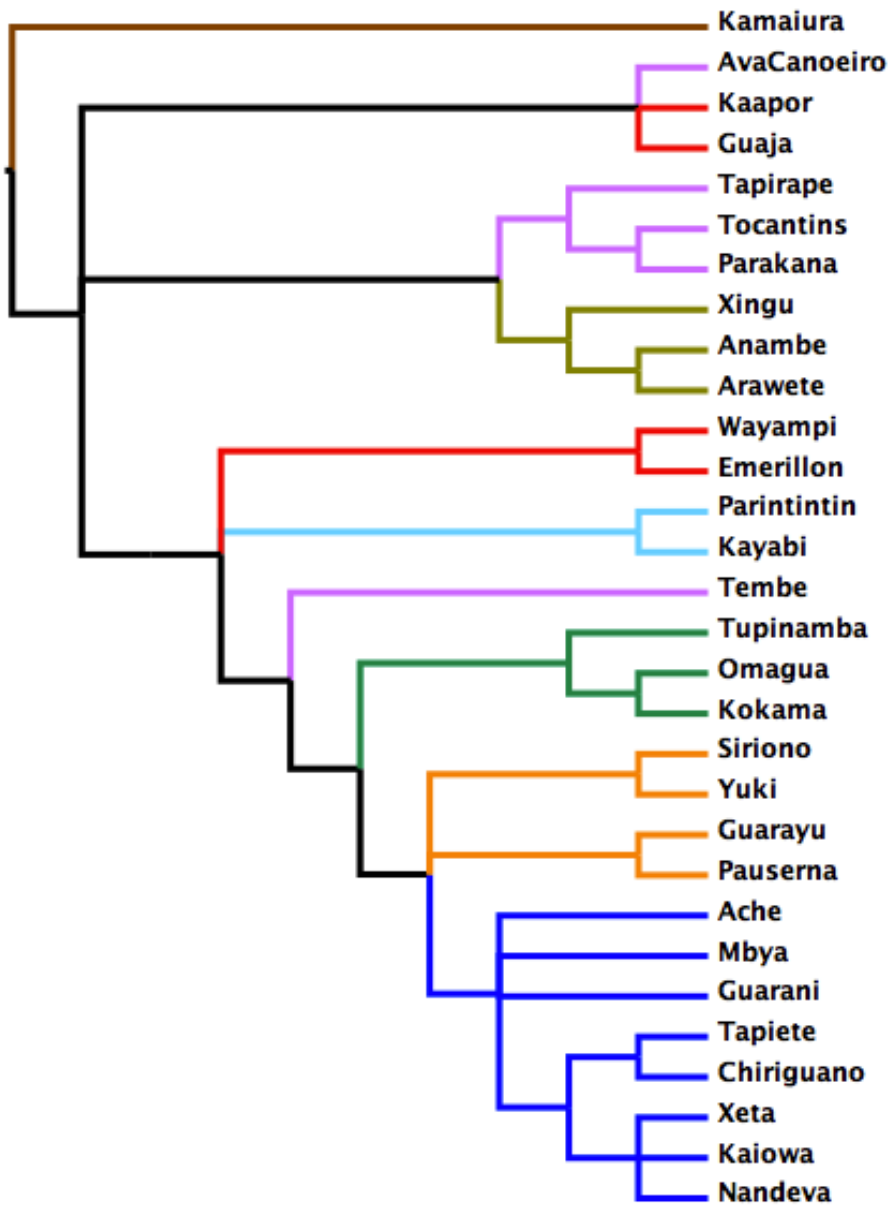
- G&A results can mostly be obtained from the QC results by eliminating higher-level structure, and merging the affected languages and subgroups into large polytomies
- They differ in that the loss of higher-level structure in the G&A tree means that well supported subgroups in the QC coding disappear

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





G&A & QC Tupí-Guaraní Classifications II

Subgroup	G&A	QC
I	X	✓
II	X	X
III	X	✓
IV	X	X
V	X	✓
VI	✓	✓
VII	✓	✓
VIII	X	X

Table 1: Subgroups Recovered by G&A and QC Codings

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

Bouckaert, Remco; Philippe Lemey; Michael Dunn; Simon J. Greenhill; Alexander V. Alekseyenko; Alexei J. Drummond; Russell D. Gray; Marc A. Suchard; and Quentin D. Atkinson. 2012. Mapping the Origins and Expansion of the Indo-European Language Family. *Science* 337(6097):957–960.

Bowern, Claire and Quentin D. Atkinson. 2012. Computational Phylogenetics and the Internal Structure of Pama-Nyungan. *Language* 88(4):817–845.

Campbell, Lyle. 1997. *American Indian Languages: The Historical Linguistics of Native America*. New York: Oxford University Press.

Corrêa da Silva, Beatriz Carretta. 2007. Mais fundamentos para a hipótese de Rodrigues (1984/1985) de um proto-awetí-tupí-guaraní. *Línguas e Culturas Tupí*, edited by Ana Suelly Arruda Câmara Cabral and Aryon Dall'Igna Rodrigues, Campinas: Editora Curt Nimuendajú, 219–239.

Corrêa da Silva, Beatriz Carretta. 2010. *Mawé/awetí/tupí-guaraní: Relações lingüísticas e implicações históricas*. PhD dissertation, Universidade de Brasília.

Drude, Sebastian. 2006. On the Position of the Awetí Language in the Tupí Family. *Guaraní y Mawetí-Tupí-Guaraní: Estudos históricos y descriptivos sobre una familia lingüística de América del Sur*, edited by Wolf Dietrich and Haralambos Symeonidis, Berlin: LIT Verlag, 11–45.

Drude, Sebastian. 2011. Awetí in Relation with Kamayurá: The Two Tupian Languages of the Upper Xingu. *Alto Xingu: Uma Sociedade Multilíngue*, edited by Bruna Franchetto, Rio de Janeiro: Museu do Índio; Fundação Nacional do Índio (FUNAI), 155–191.

References II

Forster, Peter and Alfred Toth. 2003. Toward a Phylogenetic Chronology of Ancient Gaulish, Celtic, and Indo-European. *Proceedings of the National Academy of Sciences*, vol. 100, vol. 100, 9079–9084.

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.* .

Gray, Russell D. and Quentin D. Atkinson. 2003. Language-tree divergence time support the Anatolian theory of Indo-European origin. *Nature* 426:435–439.

Gray, Russell D.; Alexei J. Drummond; and Simon J. Greenhill. 2009. Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement. *Science* 323(5913):479–483.

Greenhill, Simon J. and Russell D. Gray. 2005. Testing Population Dispersal Hypotheses: Pacific Settlement, Phylogenetic Trees and Austronesian Languages. *The Evolution of Cultural Diversity: Phylogenetic Approaches*, edited by Ruth Mace; Clare J. Holden; and Stephen Shennan, London: UCL Press, 31–52.

Greenhill, Simon J. and Russell D. Gray. 2009. Austronesian Language Phylogenies: Myths and Misconceptions about Bayesian Computational Methods. *Austronesian Historical Linguistics and Culture History: A Festschrift for Robert Blust*, edited by Alexander Adelaar and Andrew Pawley, Canberra: Pacific Linguistics, 1–23.

References III

Greenhill, Simon J.; Alexei J. Drummond; and Russell D. Gray. 2010. How accurate and robust are the phylogenetic estimates of Austronesian language relationships? *PLoS ONE* 5(3):e9573.

Jensen, Cheryl. 1999. Tupí-Guaraní. *The Amazonian Languages*, edited by R. M. W. Dixon and Alexandra Y. Aikhenvald, Cambridge: Cambridge University Press, 125–163.

Kamaiurá, Waryý. 2012. *Awetí e tupí-guaraní: Relações genéticas e contato lingüístico*. MA thesis, Universidade de Brasília.

Kaufman, Terrence. 1994. The Native Languages of South America. *Atlas of the World's Languages*, edited by C. Mosley and R.E. Asher, London/New York: Routledge, 46–76.

Kaufman, Terrence. 2007. South America. *Atlas of the World's Languages*, edited by R. E. Asher and Christopher Moseley, London/New York: Routledge, 61–93, 2 edn.

Nakhleh, Luay; Don Ringe; and Tandy Warnow. 2005. Perfect Phylogenetic Networks: A New Methodology for Reconstructing the Evolutionary History of Natural Languages. *Language* 81(2):382–420.

Ringe, Don; Tandy Warnow; and Ann Taylor. 2002. Indo-European and Computational Cladistics. *Transactions of the Philological Society* 100(1):59–129.

Rodrigues, Aryon Dall'Igna. 1984/1985. Relações internas na família lingüística tupí-guaraní. *Revista de Antropologia* 27/28:33–53.

Rodrigues, Aryon Dall'Igna. 1986. *Línguas brasileiras: Para o conhecimento das línguas indígenas*. São Paulo: Edições Loyola.

References IV

Rodrigues, Aryon Dall'Igna. 1999. Tupí. *The Amazonian Languages*, edited by R. M. W. Dixon and Alexandra Y. Aikhenvald, Cambridge: Cambridge University Press, 107–124.

Rodrigues, Aryon Dall'Igna and Ana Suelly Arruda Câmara Cabral. 2002. Revendo a classificação interna da família tupí-guaraní. *Línguas Indígenas Brasileiras: Fonologia, Gramática e História*, edited by Ana Suelly Arruda Câmara Cabral and Aryon Dall'Igna Rodrigues, Belém: Editora Universitária, Universidade Federal do Pará, 327–337.

Rodrigues, Aryon Dall'Igna and Ana Suelly Arruda Câmara Cabral. 2012. Tupian. *The Indigenous Languages of South America: A Comprehensive Guide*, Berlin: De Gruyter Mouton, 495–574.

Rodrigues, Aryon Dall'Igna and Wolf Dietrich. 1997. On the Linguistic Relationship Between Mawé and Tupí-Guaraní. *Diachronica* 14(2):265–304.

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

	GYU	TPN	TPR	PAR	KAA
stomach	ʔɨe	ɨwe	...	aw	pɨʔa
intestines	epoʔi	ɨβɨɲ	ɨe	ɨe	pɨʔa
liver	pɨʔa	pɨʔa	pɨʔã	pɨʔa	pɨʔa
G&A (Multistate)					
stomach	1	1	?	2	3
intestines	1	2	3	3	4
liver	1	1	1	1	1
G&A (Binary)					
STOMACH1 (**ɨwe*)	1	1	?	0	0
STOMACH2 (**aβ*)	0	0	?	1	0
STOMACH3 (**pɨʔ a*)	0	0	?	0	1
INTESTINES1 (**epoʔi*)	1	0	0	0	0
INTESTINES2 (**ɨβɨɲ*)	0	1	0	0	0
INTESTINES3 (**ɨwe*)	0	0	1	1	0
INTESTINES4 (**pɨʔ a*)	0	0	0	0	1
LIVER1 (**pɨʔ a*))	1	1	1	1	1
QC (Binary)					
STOMACH1 (**aβ*)	0	0	?	1	0
INTESTINES1 (**epoʔi*)	1	0	0	0	0
INTESTINES2 (**ɨβɨɲ*)	0	1	0	0	0
INTESTINES3 (**ɨwe*)	1	1	1	1	0
LIVER1 (**pɨʔ a*)	1	1	1	1	1