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

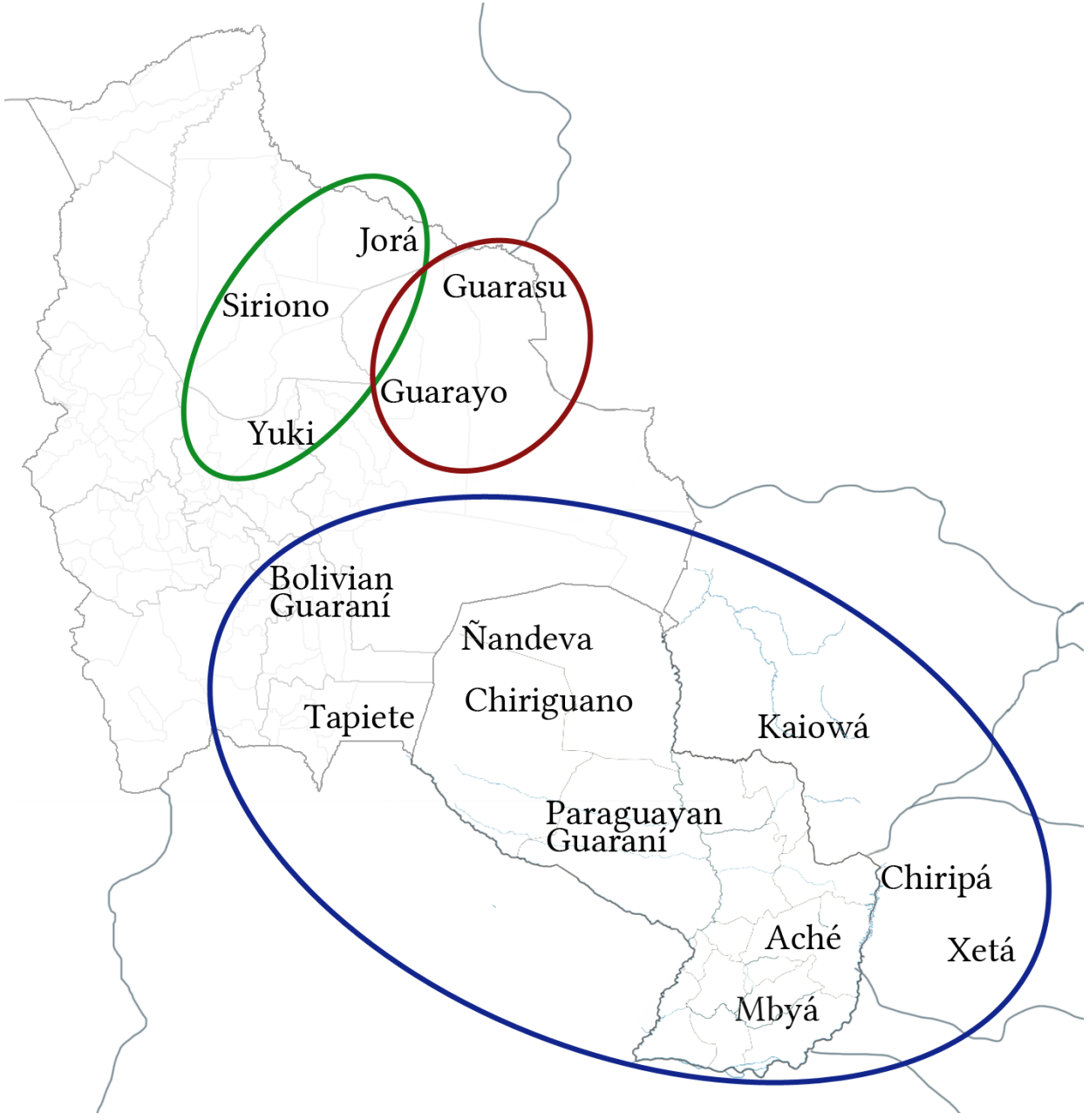
[3] University of California, Berkeley

The 18th Annual Workshop on American Indigenous Languages
May 8^{th} 2015

Southern Tupí-Guaraní languages

Two subgroups (Rodrigues and Cabral 2012)

Three clades (Michael et al. to appear)



Areal linguistics

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

A changing classification

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

Quadro I: Nova constituição interna da família Tupí-Guaraní

Ramo I:	• Guaraní Antigo • Kaiwá (Kayová, Pãi), Ñandeva (Txiripá), Guaraní Paraguaio • Mbyá • Xetá (Serra dos Dourados) • Tapieté, Chiriguano (Ava), Izoceño (Chané) • Guayakí (Aché)
Ramo II:	• Guarayo (Guarayú), • Sirionó, Horá (Jorá) • Yúki

Rodrigues and Cabral (2002:335)

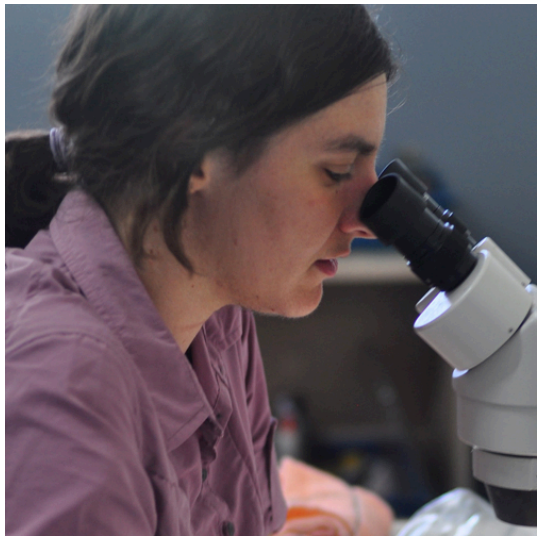
Recent changes

- 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

Tupí-Guaraní Lexical Comparative Project Team

Noé Gasparini

Swintha Danielsen



Natalia Chousou-Polydouri

Zachary O'Hagan

Keith Bartolomei

Lev Michael

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 Nakhleh 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%

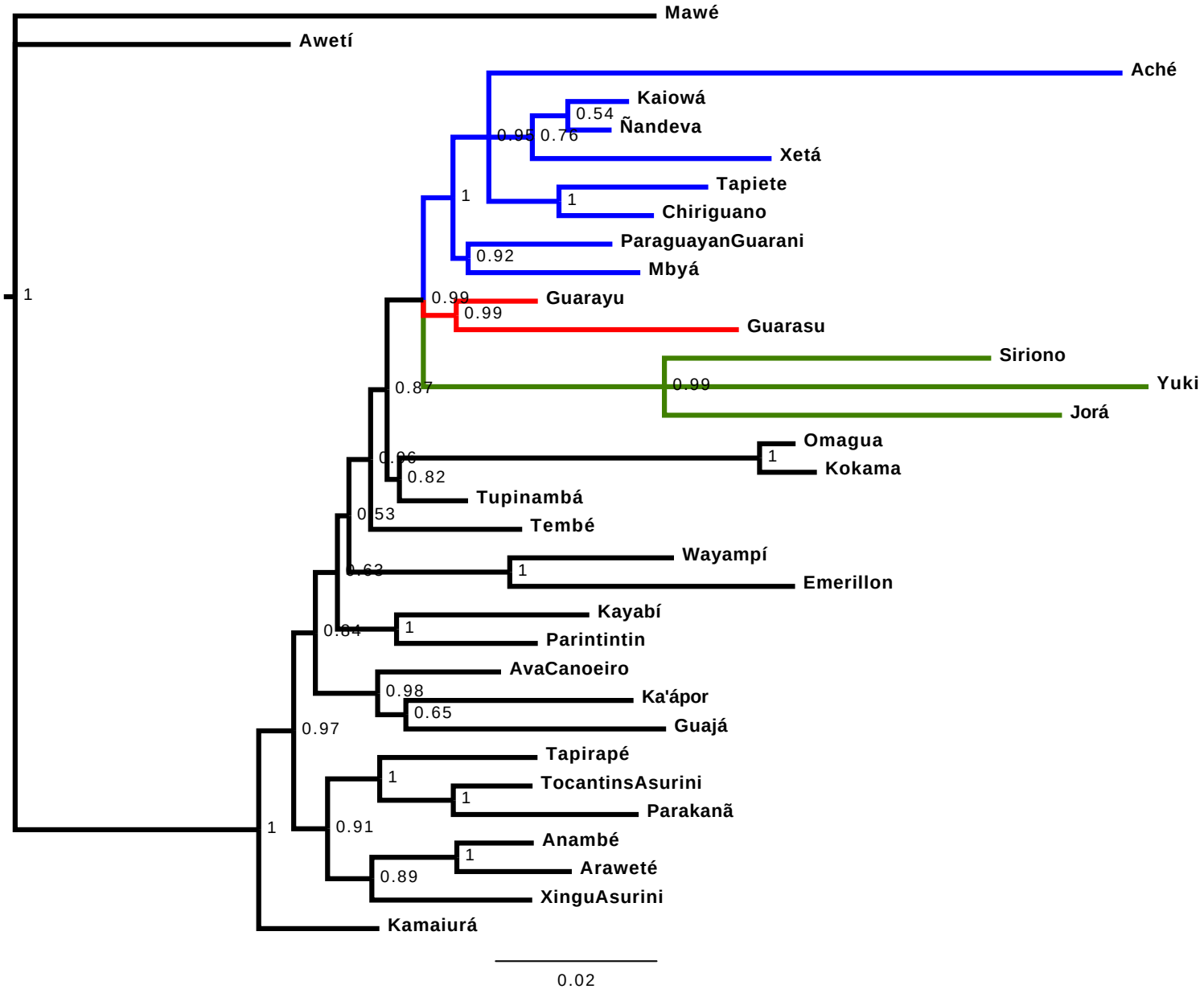
Lexical Coverage

Aché	85%	Guarayu	85%
Kaiowá	39%	Guarasu	63%
Ñandeva	20%	Siriono	92%
Xetá	32%	Yuki	82%
Tapiete	84%	Jorá	21%
Chiriguano	80%		
Paraguayan Guarani	94%		
Mbyá	84%		
Omagua	89%	Kokama	88%
Tupinambá	94%	Tembé	98%
Wayampí	89%	Emerillon	77%
Kayabí	59%	Parintintin	85%
Avá-Canoeiro	51%	Ka'ápor	83%
Guajá	45%	Tapirapé	69%
Tocantins Asuriní	83%	Parakanã	75%
Anambé	31%	Araweté	55%
Xingú Asuriní	51%	Kamaiurá	75%
Mawé	77%	Awetí	76%

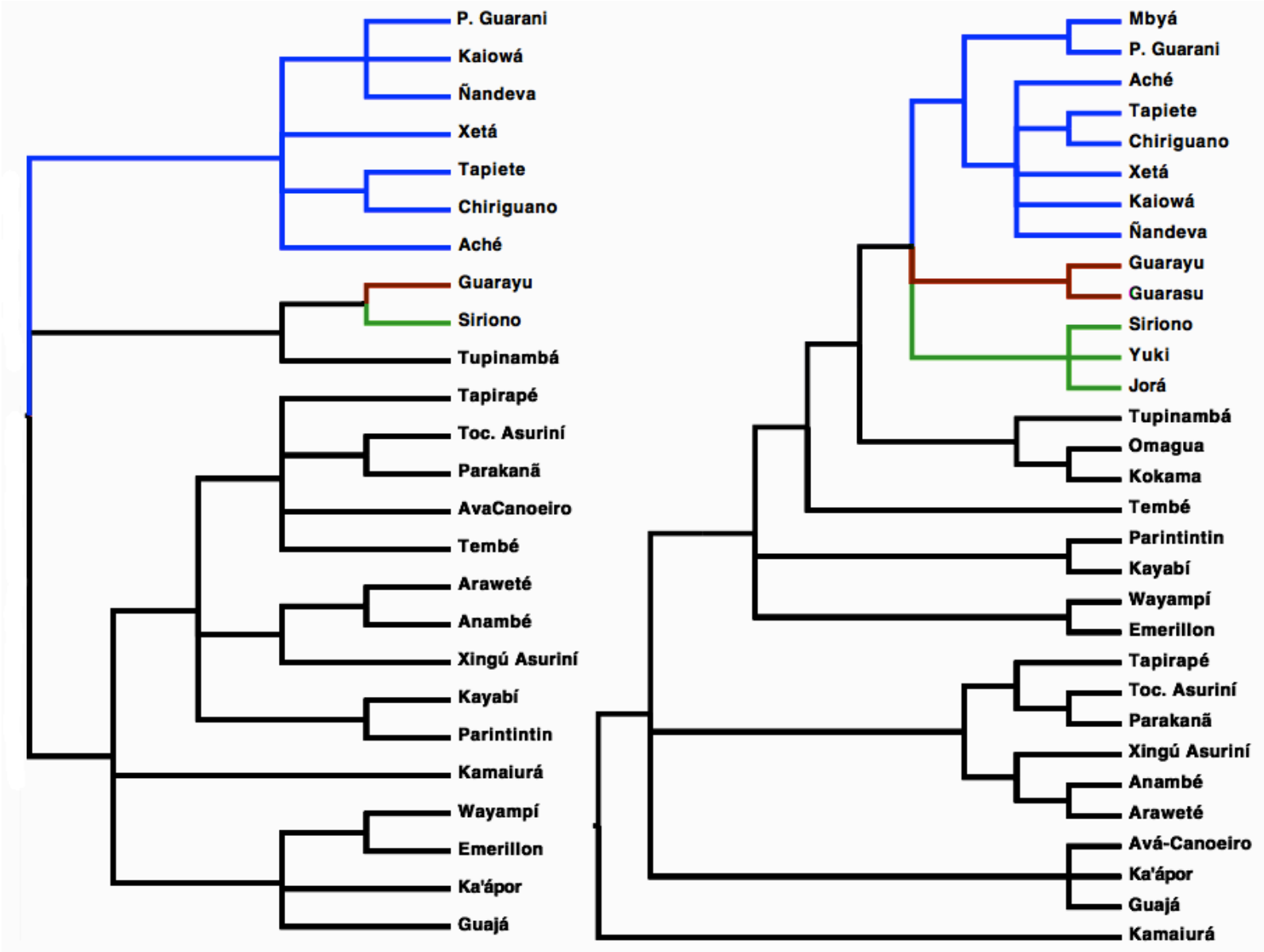
Computational Phylogenetic Methods

- 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.

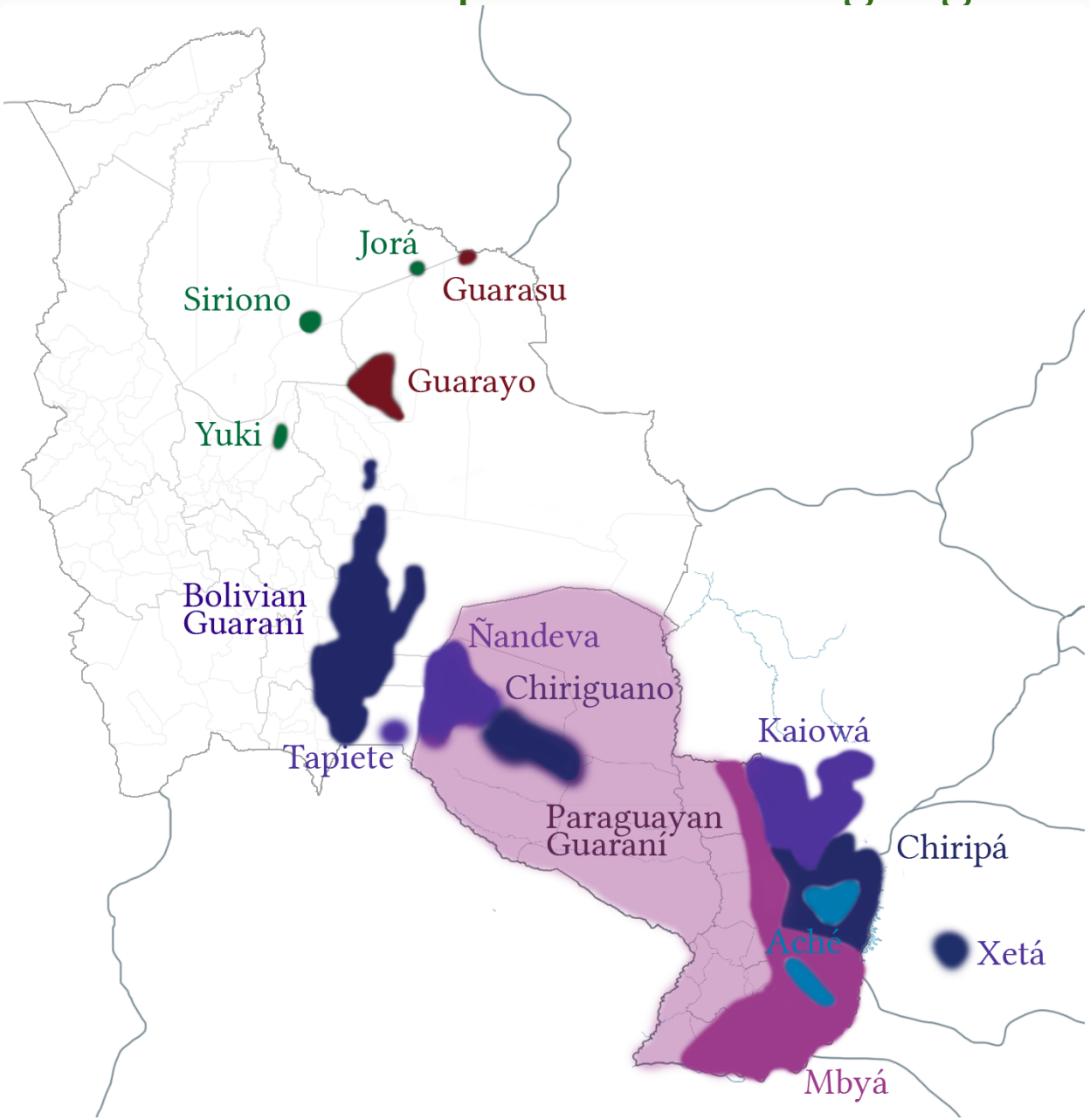
New phylogenetic classification, 2015



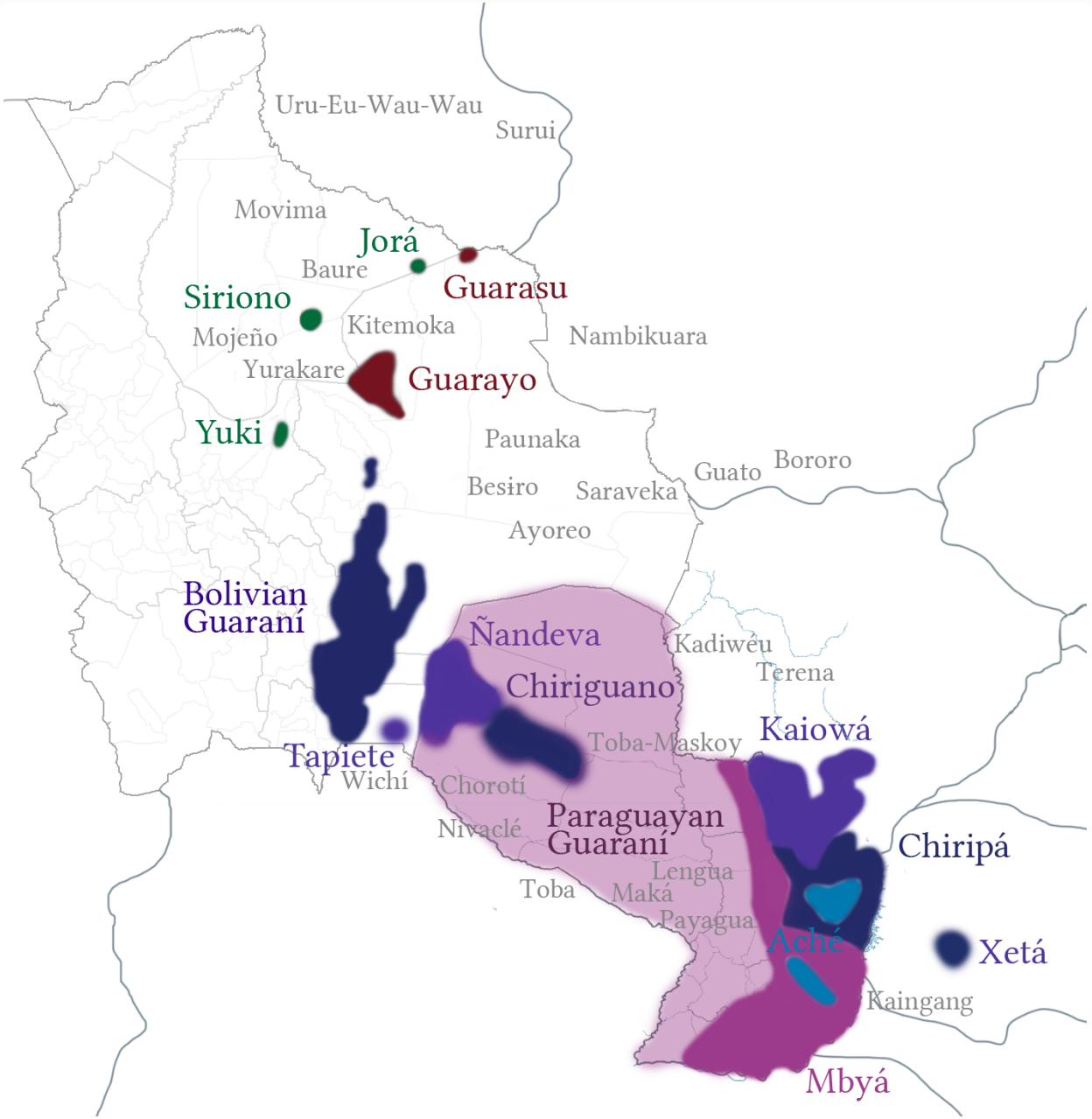
Rodrigues and Cabral (2002) and phylogenetic classification



Southern Tupí-Guaraní languages



Possible sources of loanwords



Looking for contact?

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)

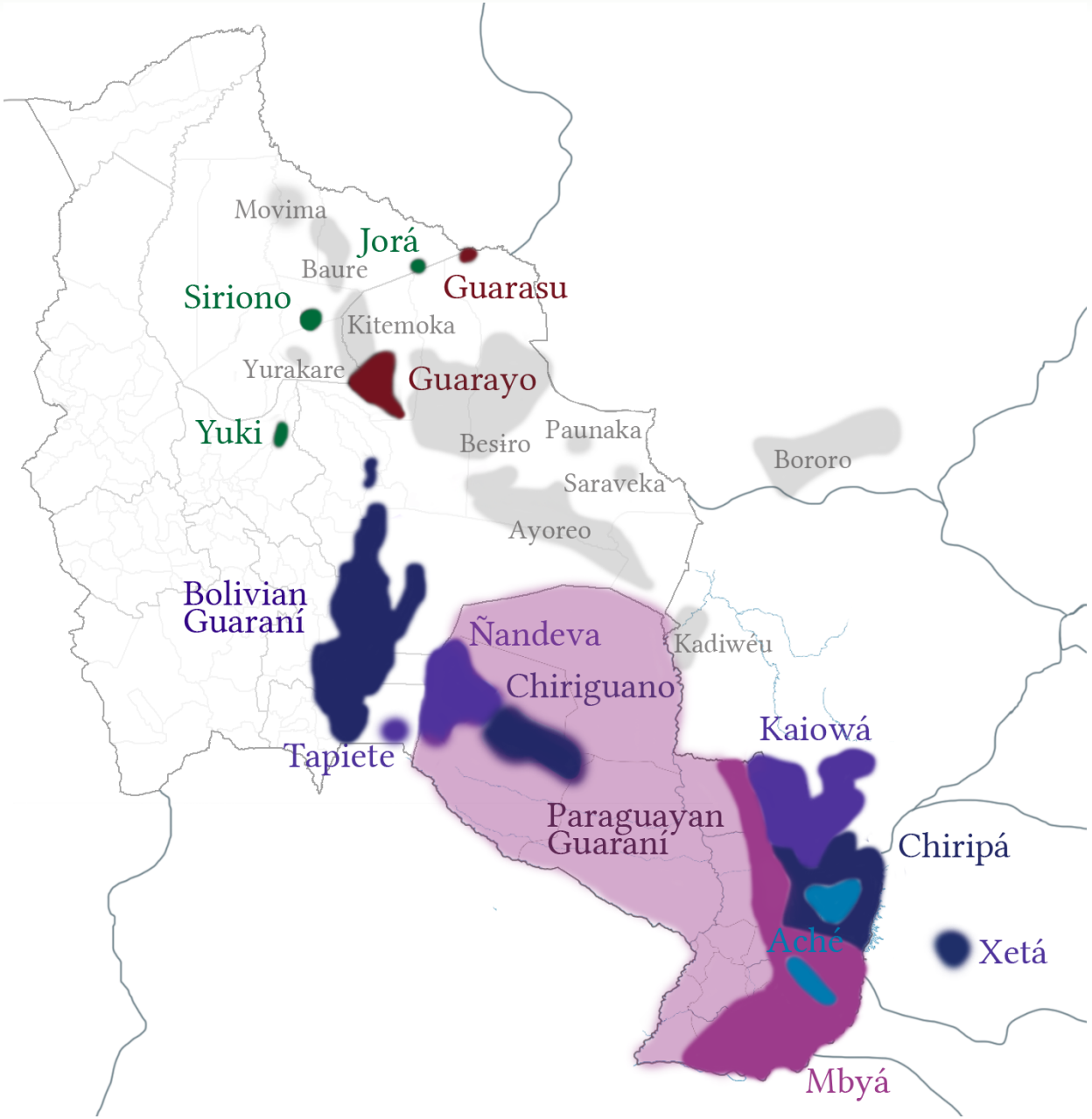
A clear split

Siriono, Yuki and Jorá show a substantial amount of differences.

Sources

Language	Family	Source
Baure	Arawakan	Danielsen database
Paunaka	Arawakan	Danielsen database
Bésɨro	Macro-Gê	Courtesy of Pierric Sans
Ayoreo	Zamucoan	Courtesy of Luca Ciucci
Kitemoka	Chapakuran	Weinold (2012)
Bororo	Bororoan	Ochoa (1995)
Kadiwéu	Guaycuruan	Griffiths (2002)
Movima	Isolate	Judy and de Judy (1962)
Saraveka	Arawakan	de Créqui-Montfort and Rivet (1913)
Yurakaré	Isolate	Ribera (1991)

Areas


Movima
Jorá
Baure
Guarasu
Siriono
Kitemoka
Yurakare
Guarayo
Yuki
Besiro
Paunaka
Bororo
Saraveka
Ayoreo
Bolivian
Guaraní
Ñandeva
Kadiwéu
Chiriguano
Kaiowá
Tapiete
Paraguayan
Guaraní
Chiripá
Aché
Xetá
Mbyá

A needle in the haystack

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:

Meaning	Sirionoan	Baure	Bésɨro	Ayoreo	Saraveka	Yurakaré	Kitemoka	Movima
Mother	**taĩʔ**	nen	piapá	daté, ité	mama-zudzani	meme	mama	ma'a
Sun	**tenda**	ses	sú	wedé			putyo	tinno
Deer	**kʲikʷã**, **jikan**	kajaw; nor	ɨbó	erámoro	tayetara	toshta	chirop	daujes
Back (n.)	**kite**	-chipi	chakú	guide-(boi)		dyuluta	totoyache	duducan
Dirt	**e-túbi**	topopkon	kuʧuki	topié			yaisaka	choesni

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 (Awetí)
  - 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 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.

Danielsen, Swintha and Noé Gasparini. Submitted. *News on the Jorá (Tupí-Guaraní) in the Bolivian Amazonia*. Boletim do Museu Paraense Emilio Goeldi, Ciencias Humanas at Belém.

de Créqui-Montfort, Georges and Paul. Rivet. 1913. Linguistique bolivienne: La langue saraveka. *Journal de la Société des Américanistes* X:497–540.

Dietrich, Wolf. 1990. *More Evidence for an Internal Classification of Tupí-Guaraní Languages*. Berlin: Gebr. Mann Verlag.

Dietrich, Wolf. 2010. O tronco tupi e as suas famílias de línguas: Classificação e esboço tipológico. *O português e o tupi no Brasil*, edited by Wolf Dietrich and Volker Noll, São Paulo: Editora Contexto, 9–26.

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.; 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.

Griffiths, Glyn. 2002. *Dicionário de língua kadiwéu*. SIL, URL `http://www.sil.org/americas/brasil/publcns/dictgram/KDDict.pdf`.

Jensen, Cheryl. 1998. Comparative Tupí-Guaraní Morphosyntax. *Handbook of Amazonian Languages*, vol. 4, edited by Desmond C. Derbyshire and Geoffrey K. Pullum, New York: Mouton de Gruyter, 489–618.

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* .

References II

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.

Ochoa, Gonçalo C. 1995. *Dicionário Bororo-Português*. Campo Grande: UCDB.

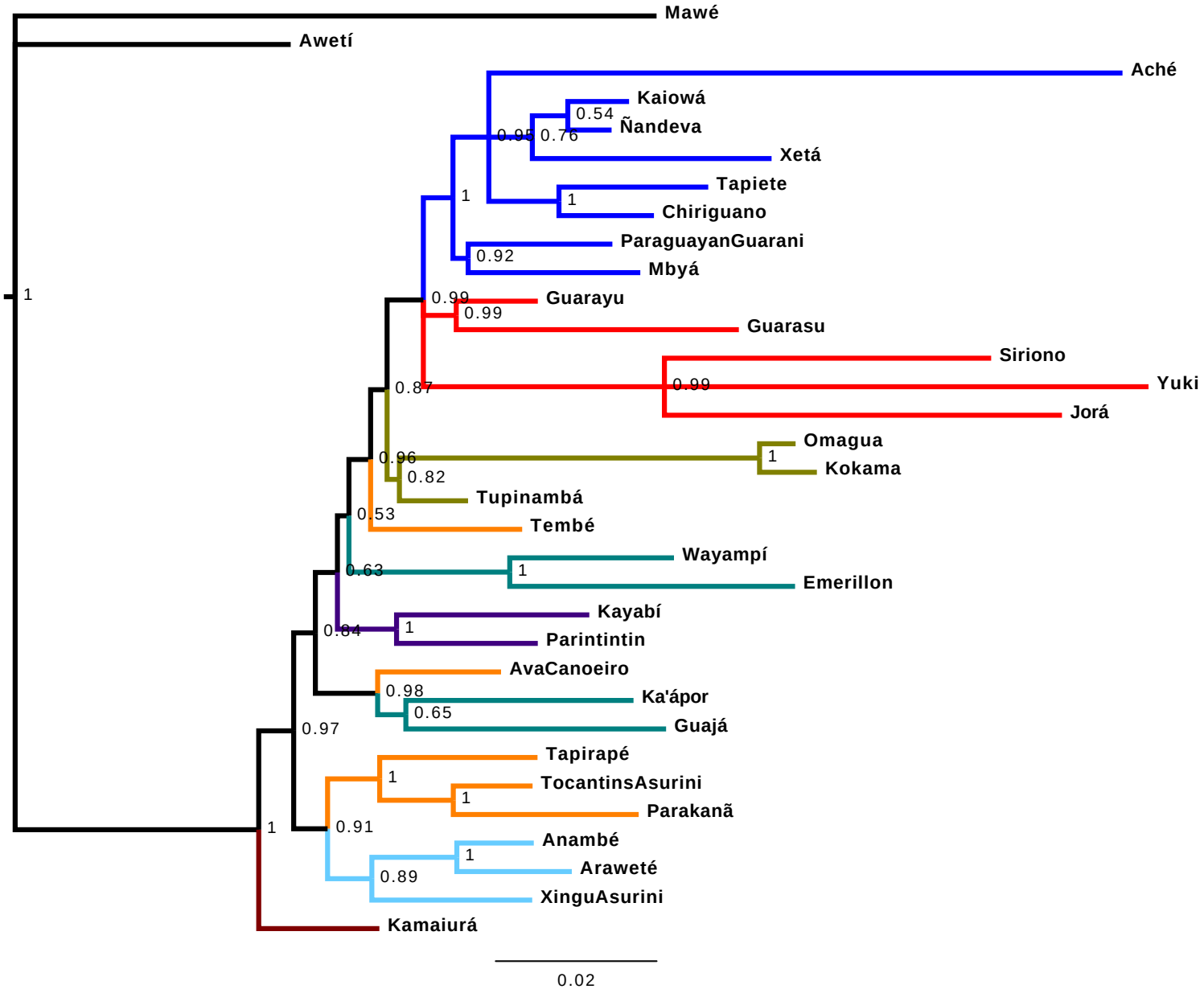
Ribera, Julio & Walter Rivero & Ascencio Rocha. 1991. *Vocabulario yuracaré-castellano, castellano-yuracaré*. Trinidad: Equipo Pastoral; MISEREOR.

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.

Weinold, Christin. 2012. *The Chapacuran Language Family: Kitemoka and Napeka, Documentational Work on Two Extinct Languages.* Master's thesis, University of Leipzig.

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