A Bayesian Phylogenetic Classification of Tupí-Guaraní

Lev Michael, Natalia Chousou-Polydouri, Keith Bartolomei
Erin Donnelly, Vivian Wauters, Sérgio Meira, Zachary O'Hagan*

Abstract

This paper presents an internal classification of Tupí-Guaraní based on a Bayesian phylogenetic analysis of lexical data from 30 Tupí-Guaraní languages and 2 non-Tupí-Guaraní Tupian languages, Awetí and Mawé. A Bayesian phylogenetic analysis using a generalized binary cognate gain and loss model was carried out on a character table based on the binary coding of cognate sets, which were formed with attention to semantic shift. The classification shows greater internal structure than previous ones, but is congruent with them in several ways.[1]

1 Introduction

This paper proposes a new classification of the Tupí-Guaraní (TG) language family based on the application of computational phylogenetic methods to lexical data from 30 TG languages

*Affiliations for the authors of this paper are: Chousou-Polydouri, Donnelly, Michael, O'Hagan—University of California, Berkeley; Meira—Museu Paraense Emílio Goeldi; Bartolomei, Wauters—Independent Scholar.

[1]We are indebted to Sebastian Drude, Françoise Rose, Eva-Maria Rößler, and Rosa Vallejos, who kindly shared unpublished lexical data from Awetí, Emerillon, Aché, and Kokama-Kokamilla, respectively. Noé Gasparini provided access to data on Yuki and Anambé. We thank Françoise Rose and audiences at **dhworom*, a UC Berkeley historical linguistics working group, the 2013 Workshop on American Indigenous Languages at the University of California, Santa Barbara, and Amazonicas V, in Belém, Brazil, for helpful comments on earlier versions of this work. Diamantis Sellis facilitated the automated binary coding of the dataset and developed scripts to verify consistency between comparative and cognate lists. This work was supported by an NSF DEL award (#0966499 *Collaborative Research: Kokama-Kokamilla (cod) and Omagua (omg): Documentation, Description, and (Non-)Genetic Relationships*) awarded to Lev Michael, and a UC Berkeley Social Science Matrix seminar grant awarded to the same author.

and two non-TG Tupian languages, Mawé and Awetí, which serve as outgroups for the phylogenetic analysis. This analysis successfully replicates many of the lower-order subgroups proposed in previous classifications (e.g., Rodrigues (1984/1985)), but yields a significantly more articulated tree structure that includes higher-order subgroups that do not emerge in any previous internal classification of the family (cf. Rodrigues and Cabral (2002)). Phylogenetic methods have been extended to the study of Austronesian,[2] Indo-European,[3] and Pama-Nyungan (Bowern and Atkinson 2012). However, with notable exceptions (Walker and Ribeiro 2011), these methods have not been applied to language families of South America.

Our results indicate that TG exhibits a relatively nested structure mainly consisting of small groups splitting off from large ones. At the highest level this is manifest in a first-order split between Kamaiurá and the rest of the family, which we call 'Nuclear Tupí-Guaraní'. Nuclear TG consists of three subgroups: a small Eastern group consisting of Avá-Canoeiro, Ka'apor, and Guajá; a medium-sized Central group consisting of two branches, 1) Tapirapé, Parakanã, and Tocantins Asuriní, and 2) Xingú Asuriní, Anambé, and Araweté; and the massive Peripheral group, consisting of the remainder. Peripheral likewise splits into three groups – two small ones consisting of Wayampí and Emerillon, and Kayabí and Parintintin – and then the large Diasporic group, consisting of the remainder.[4]

1.1 The Tupí-Guaraní Family: An Overview

The Tupí-Guaraní family includes over forty recognized varieties, with members in Brazil, Argentina, Bolivia, French Guiana, Paraguay, and Peru (Figure 1).[5] Despite its geographical extent, the time depth of TG is generally believed to be 2,000 to 3,000 years (Noelli 2008), less than that attributed to the larger Tupian stock of which it is a part (Rodrigues and Cabral 2012). Some TG varieties exhibit significant mutual intelligibility; others, in contrast, diverge radically from the typical TG grammatical profile, as in the case of Kokama-Kokamilla (Cabral 1995), Omagua (Michael 2014), Aché (Röβler 2008), and Xetá (Rodrigues 1978).

[2]Gray et al. (2009); Greenhill and Gray (2005, 2009); Greenhill et al. (2010).
[3]Bouckaert et al. (2012); Chang et al. (2015); Forster and Toth (2003); Gray and Atkinson (2003); Nakhleh et al. (2005); Ringe et al. (2002); Warnow et al. (2004).
[4]The reader is referred to §5.1 for further details.
[5]Shaded areas correspond to widespread languages. For language names and abbreviations, see Table 1.

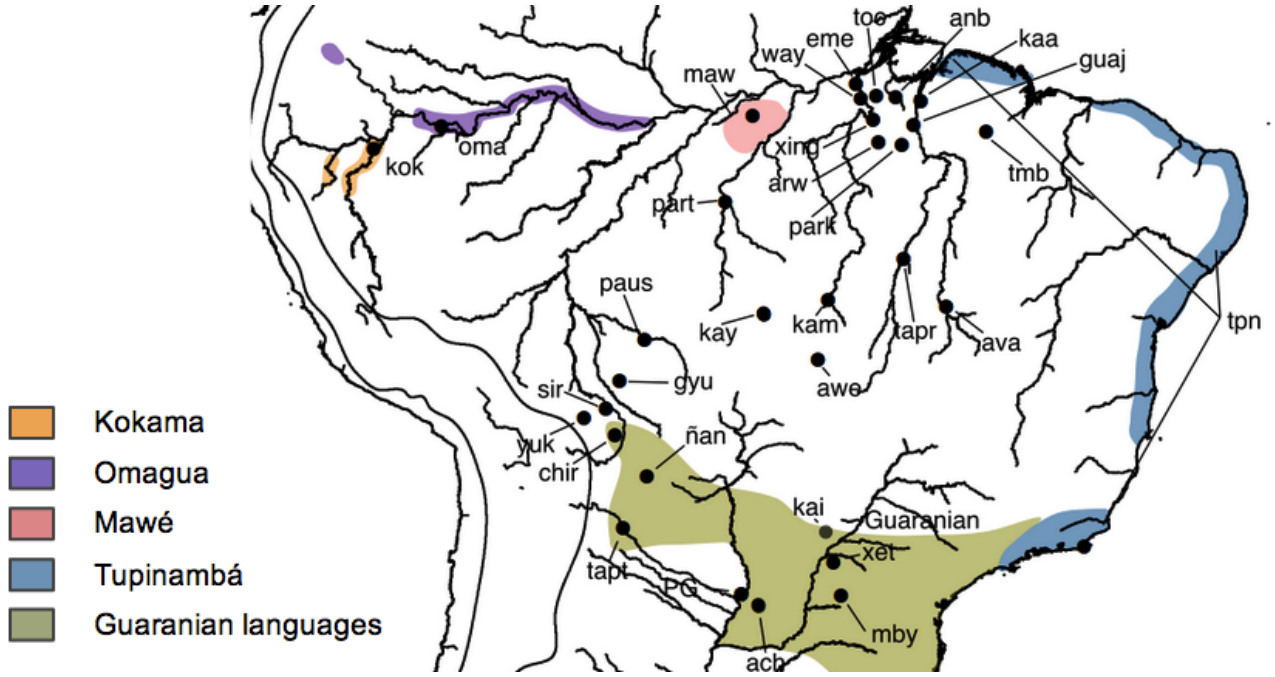

Figure 1: Earliest Known Distributions of Tupí-Guaraní Languages

The most influential classification of TG is Rodrigues (1984/1985), which is based on a combination of geographical criteria and postulated sound changes. This classification divides the family into eight subgroups, but does not propose any higher-level subgroups, yielding a rake-like structure for the family. Dietrich (1990), using quantitative distance measures on a grammatical feature dataset, argued for a more articulated structure, proposing two main branches, a southern and an Amazonian one. Rodrigues and Cabral (2002) added languages for which new data had become available, changed the position of certain languages, and proposed a subgroup corresponding roughly to Dietrich's Amazonian branch.[6]

2 Dataset

The classification presented in this paper is based on the study of lexical evolution in Tupí-Guaraní languages. As is standard in studies of this type, the empirical basis of this classification is a comparative lexical database. Data collection for this lexical database relied on harvesting words in the target Tupí-Guaraní languages corresponding to 543 meanings,

[6]Other historical works on TG include Jensen (1989, 1998, 1999), Mello (2000, 2002), Rodrigues and Dietrich (1997), and Schleicher (1998).

including numerals, body parts, plants, animals, kinship terms, natural features and phenomena, material culture items, and culturally and areally appropriate adjectives and verbs. These selected meanings were glossed in English, Spanish, Portuguese, and French to facilitate data collection from the wide variety of sources available to us. To increase the likelihood of finding cognates, near synonyms were also harvested from lexical sources.

As we discuss below, once the initial phase of cognate set construction was completed, we were able to search for cognates on the basis of their predicted forms in particular languages, using sound correspondences evident in the data. This latter search process effectively freed us from dependence on the meanings in the comparative list. Note that our original set of meanings was expanded from 500 to 543 meanings to include what emerged to be common meanings in cognate sets that were outside of our initial set of meanings (see §3).

Lexical data was collected from 30 Tupí-Guaraní languages spanning all eight conventionally recognized subgroups, as well as from two non-TG Tupian languages, Awetí and Mawé, which serve as outgroups for the phylogenetic analysis. The 30 TG languages that were selected include all languages for which we had access to non-trivial quantities of lexical data. Table 1 lists these languages, our abbreviations, and the lexical coverage for each language, that is, the percentage of meanings for which we found lexical items.

The lexical coverage in our dataset ranges from 20% (Ñandeva) to 98% (Tembé), with a mean of 71%. It should be noted that some researchers exclude low-coverage languages, using some arbitrary percentage, e.g. 50%, as a cutoff (Bowern and Atkinson 2012). However, in the biological phylogenetic literature, simulations and empirical studies have shown that, while organisms with a lot of missing data (corresponding to low-coverage languages), such as fossils, can sometimes cause the resulting phylogenetic trees to be poorly resolved, such organisms can often be accurately placed on phylogenetic trees, and can even increase tree resolution (Wiens 2003; Wiens and Morrill 2011). For this reason, we did not exclude any language based solely on its coverage percentage.[7]

We close the description of our dataset with a discussion of our choice of outgroup

[7]It is worth noting, in response to a reviewer's comment expressing concern over the effect of low coverage languages, that we did test the effect of removing low-coverage languages such as Ñandeva, Xetá, and Anambé from the analysis. We found that these exclusions tended to increase the posterior probabilities of certain nodes (i.e., increased support for certain subgroups) by mostly modest quantities, while they did not affect the tree topology.

Table 1: Languages Included in the Dataset

Language	Abbr.	%	Language	Abbr.	%
Aché	ach	85%	Ñandeva	ñan	20%
Anambé	anb	31%	Omagua	oma	89%
Araweté	arw	55%	Parakanã	park	75%
Avá-Canoeiro	ava	51%	Paraguayan Guarani	PG	94%
Awetí	awe	76%	Parintintin	part	85%
Chiriguano	chir	80%	Pauserna	paus	58%
Emerillon	eme	77%	Siriono	sir	82%
Guajá	guaj	45%	Tapiete	tapt	84%
Guarayu	gyu	86%	Tapirapé	tapr	69%
Ka'apor	kaa	83%	Tembé	tmb	98%
Kaiowá	kai	39%	Tocantins Asuriní	toc	83%
Kamaiurá	kam	75%	Tupinambá	tpn	94%
Kayabí	kay	59%	Wayampí	way	89%
Kokama	kok	89%	Xetá	xet	33%
Mawé	maw	80%	Xingú Asuriní	xing	50%
Mbyá	mby	83%	Yuki	yuk	80%

languages, and why it is necessary to include outgroup languages in the first place. We begin by noting that actual evolutionary trees are intrinsically 'rooted', in the sense that the direction in which time flows along branches of the trees is fixed by the position of each branch relative to the root node. Some methods for producing evolutionary trees automatically yield rooted trees, e.g., the Comparative Method, where rooting is achieved by a process of reconstruction, which identifies the ancestral state for each character, that is, the proto-sound or proto-form (Crowley and Bowern 2010). However, certain combinations of data types and methods, such as lexical data analyzed with phylogenetic algorithms, do not involve direct identification of ancestral states. Moreover, phylogenetic algorithms generally operate on unrooted trees to increase their speed. For both reasons, phylogenetic algorithms require an *a posteriori* rooting procedure that identifies the position of the root.

The most common technique is the outgroup method, which relies on *a priori* knowledge of an 'outgroup' language, that is, a language that is known to be closely related but not belonging to the group of languages under study (the 'ingroup'). In applying the outgroup method, the outgroup language is included in the phylogenetic analysis and then the inferred

trees are rooted where the outgroup joins the rest of the tree, since that is the point, *ex hypothesi*, at which the ingroup split from the outgroup. Ideally, multiple outgroup languages are incorporated into the analysis, including the most closely related outgroup language, and the most distantly related outgroup language is used for rooting.

We are fortunate that there is a consensus in Tupian comparative linguistics regarding the non-TG languages most closely related to the TG family. In particular, Awetí is believed to be the sister language to all TG languages, and Mawé is believed to be the sister to the Awetí-Tupí-Guaraní clade.[8] This makes these two languages ideal outgroup languages. We included both in our dataset and we rooted the trees with Mawé.

3 Cognate Sets and Character Coding

Phylogenetic analyses are based on the comparison of homologous features, that is, features that are descended from a feature found in a common ancestor. Given the fact that our comparative data is lexical in nature, and inspired by the goal of modeling lexical evolution as closely as possible on the Comparative Method, we chose cognate sets as the basis of the homologous features used in the analysis. In particular, we selected our characters to be the presence or absence of a form belonging to a particular cognate set in a particular language. Whether a language exhibits, or fails to exhibit, a reflex of a given proto-form is a heritable feature that is largely independent of other characters in this scheme, making it a good character for phylogenetic purposes.

This character coding strategy requires that we code whether a language exhibits a form belonging to each cognate set extracted from the data. It will be readily appreciated that a challenge in implementing this coding is knowing whether a language in fact lacks a form belonging to a given cognate set or whether such a form is simply missing from the lexical resources available on that language, and thus present, but unlocatable. Below we discuss how we sought to be confident that a language lacked a form belonging to a particular cognate set, though it was, of course, impossible to be entirely certain regarding these absences.

After the initial phase of data collection, we constructed cognate sets which, crucially,

[8]See Corrêa da Silva (2007, 2010); Drude (2006, 2011); Kamaiurá (2012); Rodrigues and Dietrich (1997).

included forms that exhibited semantic shift. Forms which have undergone semantic shift still remain cognate, of course, and identifying these items is crucial for replacing bogus absences in the character table with presences. Each cognate set was labeled with its 'central' meaning,[9] and two additional data collection processes were carried out to find cognates that we might have missed in our original data collection, to which we now turn.

First, in cases where we had not yet found a form belonging to a particular cognate set in a particular language, we used sound correspondences inferred from the dataset, or already identified in published sources (e.g., Soares and Leite (1991)), to predict expected forms for the potentially present cognate, and searched for those predicted forms. This technique was particularly effective in finding missing cognates in languages with relatively extensive lexical resources (e.g., Tupinambá).

Second, we engaged in another round of data collection to systematically harvest data corresponding to central meanings that were not in our original set of meanings, but which emerged from the process of cognate set construction. In general, these additional central meanings were ones that were deducible from sets partially populated by forms with meanings that **were** included in the original set of meanings, but which had, it could be inferred, shifted from a meaning that was **not** included in the original set of meanings. When it was possible to infer what these "missing" central meanings were, we added these meanings to the comparative list and searched systematically for these meanings, thereby filling out the partially populated cognate sets in question.

For example, we found forms cognate to the root *aʔaŋ* in several languages, but with a variety of disparate meanings, including 'sing' in Tapiete and Kayabí, 'draw' in Guarayu and Tembé. Similarly, nominalized forms of the same root meaning 'spirit' were attested in Tupinambá and Paraguayan Guaraní. On the basis of these meanings, we deduced that the central meaning for the cognate set approximated 'imitate', and thus we selected 'imitate' as a new meaning to include in the comparative list. Upon searching for this new meaning, we found cognates that we had previously overlooked in most languages. We added 43 central meanings to our comparative list by this process.

[9]The 'central' meaning of a cognate set is the meaning from which all other meanings found in the cognate set could be most easily and plausibly derived. The 'central' meaning is not necessarily the most common meaning, nor does it necessarily constitute a claim regarding the root's proto-meaning.

Both of the above procedures identified additional cognates that definitively replaced possible absences in the character matrix. The remaining possible absences were either coded as true absences, if particular criteria were met, or as 'unknown', if these criteria were not met. Note that the phylogenetic algorithm treats unknowns as either present or absent, effectively removing them as a factor in distinguishing between possible trees. The criteria in question are intended to distinguish true absences from mere empirical gaps in the resources, and include: 1) that a form corresponding to the central meaning of the cognate set is in fact attested for a different cognate set for the language in question;[10] 2) no cognate was found when searching for the central meaning or near-synonymous meanings; 3) there was no compound or otherwise complex word in our dataset that incorporated the cognate for that particular language (see below); and 4) in the cases of well documented languages, no cognate was found when searching based on the expected form. If any of these criteria was not met, the possible absence was instead coded as 'unknown'.

Other aspects of the character coding that require comment are our treatment of compounds and morphologically complex forms, and our treatment of loans. During the construction of cognate sets, we encountered a large number of compounds and morphologically complex words, which we will henceforth refer to as 'compounds', since we treated them in the same way. Compounds raise two issues with respect to character coding: attestation of roots and compounding itself as a phylogenetic character. With respect to the first issue, it is clear that a compound may serve as an attestation of a root that participates in a given cognate set. If a compound did so, the relevant cognate was coded as present.[11]

Second, compounds themselves can be homologous features, as seen by considering 'star', which is typically a compound of the words 'moon' and 'fire' in TG languages. This rather unusual compound is unlikely, we suggest, to have been independently innovated more than once, and as such, the compound itself was inherited into daughter languages from the proto-language in which it was innovated. We call instances of compounding or derivation

[10]This criterion is intended to identify empirical gaps, and to not count such gaps, incorrectly, as absences (in the character sense outlined above). Clearly, if no form is found that corresponds to a given meaning, it is entirely possible that the lexical resources are simply missing this meaning, so that it would be rash to count any gaps in cognate sets with this central meaning as true absences.

[11]For example, a language may not exhibit a reflex of the root *peʔir* 'sweep' as a productive verb root, but may exhibit a word for 'broom' that was derived from *peʔir*; under such circumstances the derived word was taken as evidence for the presence of the root *peʔir* in the language.

that plausibly occurred once in the past and were then inherited as a unit in the daughter languages 'genetic compounds' and treat them as characters. Compounds that do not meet this criterion are called 'potentially independent' and not treated as characters. Note that we only consider a compound a member of a compound cognate set if: 1) all its subconstituents are cognate with the subconstituents in the other members of the compound cognate set; and 2) are in the same linear order.

We consider compounds to be genetic if any of the following criteria are met: 1) the meaning of the compound is unpredictable (see above); 2) the meaning of the compound is predictable, but the reflexes of the compound show evidence of phonological erosion (e.g., *wirapar* 'bow', instead of the uneroded *iwirapar*); 3) the meanings of the compound's constituents cannot be identified and the compound is widely distributed in many languages of our dataset; and 4) it is a singleton, i.e., found in only one language.

Turning to loans, we note that we were able to identify loans from other language families (e.g., Quechuan and Romance) into Tupí-Guaraní, as well as some from Tupí-Guaraní languages into Mawé. Loans were coded as 'singleton' (apomorphic) characters, i.e., characters that were coded as present only for one language. For example, Ka'apor, Tapiete, Omagua, and Kokama have all borrowed the word for 'mother' from Spanish or Portuguese. As all these borrowings represent independent events, there are four different apomorphic characters in the semantic group 'mother', one for each language. Although apomorphic characters are uninformative in parsimony analyses, in a likelihood and Bayesian framework, apomorphic characters are informative for the estimation of evolutionary rates and branch lengths and should not be excluded. We ultimately constructed a total of 4205 cognate sets, of which 1113 were parsimony-informative and 2989 were singleton cognate sets.

4 Phylogenetic Analysis

Our proposed classification of Tupí-Guaraní is based on Bayesian phylogenetic methods, originally developed to infer evolutionary trees for biological organisms and implemented in MrBayes3.2 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). In order to understand the utility of these methods, it is important to realize that they differ fundamen-

tally from distance-based methods, such as lexicostatistics, in two important ways. First, unlike distance-based methods, they do not seek to measure overall similarity between languages by collapsing the entirety of the comparative dataset into pairwise distances between languages, but instead seek to account for the distribution of individual character states across the dataset. In particular, phylogenetic methods evaluate a massive number of possible trees, and the evolution of characters on those trees, against the attested distribution of character values, using an optimization criterion (e.g., parsimony, which prefers trees with a smaller number of independent innovations of a character) to identify the best trees. With respect to the lexical data we examine in this paper, the phylogenetic algorithm searches the mathematical space of possible trees and looks for the tree, and the associated processes of cognate gain and loss (i.e., character evolution), that best explains the distribution of all the cognate sets we have developed, according to the optimization criteria of the model we employ (see below). As a result, like the Comparative Method, our phylogenetic analysis is capable of differentiating between shared innovations and shared retentions, and only takes the former into account for subgrouping. A partial list of the shared lexical innovations defining particular subgroups is given in Appendix A.

In order to understand how Bayesian phylogenetic inference is implemented, it is helpful to observe that these methods are a special case of the more general Bayesian approach to model evaluation, which seeks to answer the question: what is the probability that a particular model of a given phenomenon is correct, given the data relating to that phenomenon? In our case, this amounts to the question: what is the probability that a given TG tree is correct, given the cognate sets we have constructed? As such, the outcome of Bayesian model evaluation is a 'posterior probability', a value for the probability that the model is correct, after the data have been taken into account.[12] This posterior probability, $P(model|data)$, is calculated using Bayes theorem, given in (4).

(1)
$$P(model|data) = P(data|model)\frac{P(model)}{P(data)}$$

[12]In phylogenetic inference, the model includes a variety of parameters: the tree topology, the branch lengths, the transition rate matrix, the stationary probabilities, rate variation among sites parameters, etc. All these parameters at the end of the analysis have an associated 'posterior probability distribution', i.e., a distribution which shows the posterior probability over a range of values.

We now explain each expression to the right of the equality. $P(data|model)$ is the likelihood of the data under the model, or simply 'likelihood', and corresponds to the probability that the observed data (the character matrix derived from the cognate sets) would be produced by the hypothesized model (the tree and its associated parameters). The prior probability of the model, $P(model)$, consists of the various preliminary estimates for the model parameters, or 'prior distributions', with which we furnish the analysis (see below). Finally, the denominator of the formula, $P(data)$, is the probability of the data integrated over all possible parameter values, and is, in fact, impossible to calculate exactly for a phylogenetic analysis. It can, however, be approximated to an arbitrary degree of accuracy using a number of stochastic methods, with the standard in phylogenetic analysis being the Metropolis-Hastings algorithm, a Markov Chain Monte Carlo (MCMC) method.

4.1 Evolutionary Model and Prior Distributions

Since Bayesian phylogenetic inference is a model-based method, it requires both a specification of a model of evolution for the characters used in the analysis, and a choice of prior distributions (or simply 'priors') for all the parameters of the model. The most commonly used models for lexical evolution in linguistics are the generalized binary model[13] and the stochastic Dollo model (Alekseyenko et al. 2008).[14] Both models simulate the evolution of binary characters, and both are thus suitable for modeling characters based on the presence or absence of particular cognates. The main difference between the two models is that the Dollo model presupposes that each cognate originates only once on a tree, while the generalized binary model allows cognates to originate more than once.

For our analysis, we adopt a generalized binary model, since the Dollo model's single origin assumption is far too strong to be a realistic model for our data. For the Dollo model to be valid for our data, it would be necessary to identify and remove all intra-family loans, identify and remove all inter-family loans that occur more than once in the family, and correctly identify all reflexes of a given protoform. The generalized binary model, on

[13] In MrBayes, the phylogenetic software used here, this type of model is called a 'restriction site model'.

[14] The latter is implemented in BEAST (Drummond and Rambaut 2007; Drummond et al. 2012), a phylogenetic application that has been used in several phylogenetic linguistic studies.

the other hand, does not impose a unique origin of each cognate set, which accommodates potential borrowings, as well as instances in which inadequate language documentation leads us to mistakenly code a cognate as absent, in the sense that false absences like this may result in an analysis that posits two independent cognate set gain events. The generalized binary model also permits different rates of change between the two states represented in the character matrix, i.e., different rates of cognate gain and loss, which is very important for modeling cognate evolution, since independent gains of the same cognate are expected to be very rare,[15] while independent losses of a certain cognate are comparatively common.

Priors represent our expectations for the values of each parameter **before** we look at any of the data, and can be based on intuitions and prior analyses (but of different datasets, so as to avoid circularity). In cases where prior information or strong expectations are lacking, one can use uninformative or 'flat' priors that place more or less the same probability across many different values for each parameter. Bayesian inference allows the data to override (or "swamp") the prior if there is overwhelming evidence in a different direction.

The binary model has only one free parameter, the stationary probabilities of the states, which are proportional to the rates of gain and loss of cognates.[16] We selected a flat Dirichlet prior for the stationary probabilities, which gives equal probability to all possible ratios of cognate gain and loss. This prior is uninformative regarding the relative rates of cognate gain and loss, so that any asymmetry that emerges in these rates is generated by the data.

Since our dataset exhibits domains with different cognate gain and loss rates, we compared a model including gamma-distributed rate variation across cognate sets, which allows for variability in the rates of evolution across different cognate sets, to a model without rate heterogeneity, using Bayes factors. To estimate the Bayes factors we used the AICM procedure, as implemented in Tracer v1.6 (Baele et al. 2012). For the shape parameter of the gamma distribution we used a uniform prior in the interval (0, 200).

[15] In practice, independent gains would be instances of borrowing or mistaken cognacy decisions.

[16] The stationary probabilities are the proportion of each state in a number of cognate sets when they have evolved for an infinite amount of time.

4.2 MCMC and Summary of Results

MrBayes uses the Metropolis-Hastings algorithm, a Markov Chain Monte Carlo (MCMC) method, to evaluate and optimize the model parameters, including tree topologies and branch lengths. This method aims to explore a 'tree space' whose points consist of combinations of topologies, branch lengths, and parameter values, seeking out regions of high posterior probability. It does so using a set of MCMC 'chains', which are quasi-random walks through the tree space that are constructed by starting at a given point in the tree space, and then moving to random nearby points in successive iterations, called 'generations'. The chain is sampled at regular intervals, and the likelihood – as well as the combination of topology, branch lengths, and parameter values – is recorded. If the sampled posterior probability is higher than the previously sampled one, the iteration process "accepts" the new point and the chain continues from there. If it is lower, the chain may accept the new point, with a probability equal to the ratio of posterior probabilities at the lower and higher probability points. If the new point is rejected, then the chain "backtracks" to the previous point and is iterated from there. This behavior ensures that different regions in the tree space are represented in proportion to their posterior probability, with high posterior probability regions represented with greater frequency than low posterior probability ones.

For a Bayesian analysis to be trustworthy, the MCMC needs to run long enough for the posterior probability distribution to be adequately approximated with the collected samples. We infer that MCMC has run long enough when the chains have reached 'stationarity' (i.e., they mostly remain in one region of tree space), and when independent chains have 'converged' (i.e., they are all sampling from the same distribution). Because chains start at arbitrary points, the initial portion of the chain trajectory and the associated samples are not representative of the posterior probability distribution and are discarded as 'burn-in'.

The usual way to summarize the results of an MCMC run is to integrate the posterior probability of every parameter over all the possible values of the other parameters. For parameters with numerical values, we calculate estimated values and 95% credibility intervals. For tree topologies, a common summary is the construction of a majority-rule consensus tree from the sample, that is, integrating how often a subgroup is found in the sampled trees.

All analyses were performed with MrBayes3.2 at the California Academy of Sciences CCG PhyloCluster. For every analysis, we ran two independent chains of 10,000,000 generations each, logging results every 1,000 generations; we ran six 'hot' chains (three for every 'cold' one) with swaps being proposed every 50 generations.[17] We used a conservative 25% burn-in for MCMC diagnostics and our results. Stationarity and convergence for all parameters were verified using Tracer (Rambaut and Drummond 2007), while topology convergence was assessed with the average standard deviation of split frequencies, which in all cases fell below 0.01. Majority-rule consensus trees of the posterior sample were made with MrBayes3.2 and annotated with FigTree.[18] All characters were reconstructed using maximum likelihood and the estimated model parameters of cognate gain and loss on the majority-rule consensus tree with Mesquite (Maddison and Maddison 2007).

5 Results and Discussion

The majority-rule consensus tree resulting from our analysis is presented in Figure 2, with the posterior probabilities given for each node. Bayes factors (BF) comparison between runs with and without gamma-distributed rate variation across cognate sets yielded decisive support for the inclusion of rate heterogeneity (BF difference 1316 in favor of gamma-distributed rates) (Kass and Raftery 1995). The asymmetry between cognate loss and gain is 31:1, which suggests Dollo-like behavior and a low level of borrowing within the family. Appendix A lists some cognate gains and losses reconstructed at some of the well supported nodes.

5.1 Proposed Classification of Tupí-Guaraní

The classification of Tupí-Guaraní that emerges from our analysis is shown in Figure 3, which is labeled with proposed names for the well supported subgroups. Here we show only the nodes with posterior probabilities ≥ 0.80, which means that the subgroups dominated by

[17] This is an additional technique to avoid being trapped in local maxima. In this approach, parallel to the sampling chain (= cold chain), the algorithm runs a number of exploratory hot chains, which have a higher probability to move into regions of low posterior probability, thus traversing the tree space more easily. At regular intervals the states of the chains are compared and if a hot chain is at a region of higher posterior probability than the cold chain, then the hot chain in question becomes the cold chain.

[18] See: http://tree.bio.ed.ac.uk/software/figtree/.

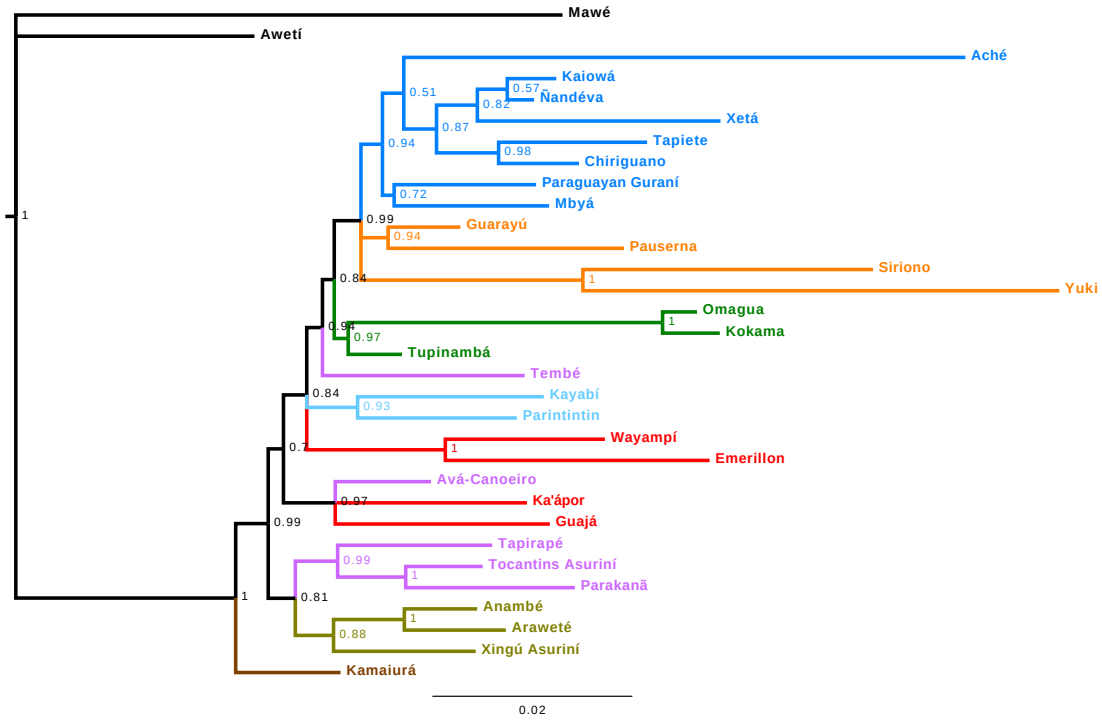

Figure 2: Majority-rule Consensus Tree with Shading from Rodrigues and Cabral (2002)

these nodes are supported by a minimum of 80% of the trees sampled by the algorithm, a cutoff we consider conservative. Nodes with lower posterior probabilities are not considered to delimit well supported subgroups, and the languages below such nodes are merged as 'polytomies' – i.e., unarticulated sets of languages – into the next higher well supported node. Before discussing the structure of this classification, it is worth noting that TG itself is recovered as a well supported subgroup ($p = 1$), confirming that it is a subgroup of Tupian.

Nuclear Tupí-Guaraní and Kamaiurá The phylogenetic analysis indicates that the highest level structure of TG family involves a two-way split, in which Kamaiurá emerges as the sister language to all other Tupí-Guaraní languages, which together comprise a single well supported subgroup ($p = 0.99$) that we call 'Nuclear Tupí-Guaraní'. This result is strikingly different from previous classifications, although it is somewhat reminiscent of Rodrigues' (1984/1985) classification of Kamaiurá as being the sole member of Group VII, an indication that it lacks close relatives. Both Lemle (1971) and Rodrigues and Cabral (2002), however, classify Kamaiurá as a member of larger subgroups.

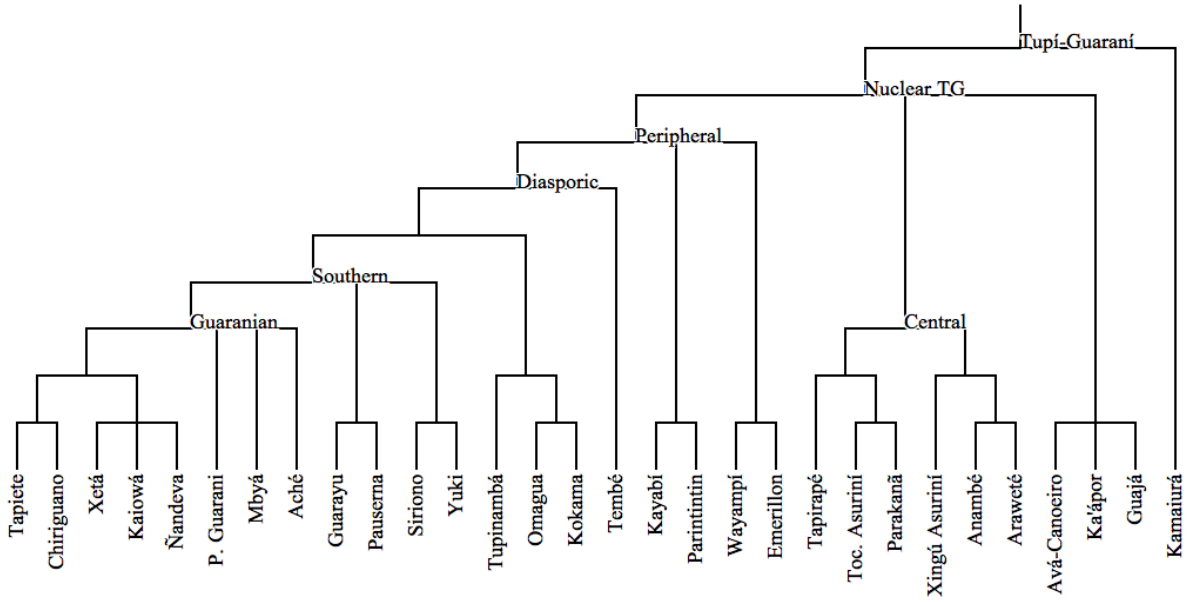

Figure 3: Proposed Classification of Tupí-Guaraní

The position of Kamaiurá in this analysis naturally raises suspicions that this result may have been influenced by loans into Kamaiurá from Awetí, which is in contact with Kamaiurá (Drude 2011), and serves as the immediate outgroup language for TG as a whole. Were there a sufficient number of Awetí loans in Kamaiurá, the phylogenetic analysis could be misled to consider Kamaiurá 'less TG-like' than the other TG languages, which could produce the resulting top-level split between Kamaiurá and Nuclear TG.

However, inspection of the cognate sets that can be reconstructed at the Nuclear TG node showed a large number of cognate gains (at least 13) and a relatively small number of cognate losses (3). If indeed the position of Kamaiurá was the effect of borrowing, we would expect many more cognate losses reconstructed at the Nuclear TG node (since the "loss" of the loans would be optimized at the Nuclear TG node) and correspondingly few cognate gains. More Kamaiurá data, and especially verification that these apparently missing Nuclear TG cognates are indeed absent, would strengthen our hypothesis and rule out the possibility that this pattern is driven by borrowing. Nevertheless, the top-level split between Kamaiurá and Nuclear TG is well supported by the current dataset.

Structure of Nuclear TG The analysis then indicates that Nuclear TG splits into three subgroups: the medium-sized Central ($p = 0.81$) subgroup, the small Eastern ($p = 0.97$) subgroup, and the large Peripheral subgroup ($p = 0.88$).

The Central subgroup includes the entirety of Rodrigues' (1984/1985) Group V plus a subset of Group VI, with most of the languages located in the Xingú-Tocantins interfluvium. Curiously, almost all the changes reconstructed for the Central subgroup are cognate losses. Central exhibits two major subgroups: one consisting of Tapirapé, Tocantins Asuriní, and Parakanã ($p = 0.99$); and another consisting of Xingú Asuriní, Araweté, and Anambé. The Eastern subgroup consists of Avá-Canoeiro, Ka'apor and Guajá, which were historically found in the Tocantins basin (Balée 1994).

Structure of Peripheral The Peripheral subgroup exhibits three well supported subgroups: one consisting of Wayampí and Emerillón ($p = 1$); another consisting of Kayabí and Parintintin ($p = 0.98$); and Diasporic (p=0.94), which includes the remaining languages. Within Diasporic, Tembé is the sister to the remaining languages, which constitute a well supported subgroup ($p = 0.83$), which itself splits into two groups, one consisting of Tupinambá, Omagua, and Kokama ($p = 0.97$), and another that we call Southern ($p = 0.98$). Tembé is the only Diasporic language located close to the main concentration of Tupí-Guaraní languages near the mouth of the Amazon, with all other Diasporic languages located at the edges of the Amazon basin, or outside it. The Diasporic subgroup is supported by at least 9 cognate gains (see Appendix A).

Southern Subgroup The Southern subgroup exhibits a three-way split into the Siriono-Yuki subgroup ($p = 1$), the Guarayu-Pauserna subgroup ($p = 0.95$), and the large Guaranian subgroup ($p = 0.92$). This trichotomy may be resolvable with additional, or different kinds of data, or it may be an indication of rapid differentiation of Proto-Southern in these three branches. The Siriono-Yuki and the Guarayu-Pauserna subgroups are located in the vicinity of the headwaters of the Madeira River, while the Guaranian subgroup is spread across the Paraná basin. The Southern subgroup is one of the better supported subgroups in our analysis, with 7 cognate gains and 11 cognate losses reconstructed at this level.

Guaranian Subgroup The Guaranian languages form a well supported subgroup, but its internal structure is not well resolved. This is not entirely surprising, since several of the varieties comprising the group exhibit significant mutual intelligibility. Two notable languages within Guaranian are Aché and Xetá. Both exhibit grammatical features which diverge significantly from the typical TG profile, and have been thought to have been significantly affected by language contact (Rodrigues 1978; Röβler 2008). Both Aché and Xetá show instability in terms of where they attach on the tree and are at the tips of some of the longest branches, which indicates significant lexical change, and show that there is a correlation between the amount of grammatical and lexical change in these languages. The Guaranian subgroup is supported by 6 cognate gains and 4 losses.

5.2 Comparison with Previous Classifications

In this section we compare our classification with the major previous classifications of the family, including Lemle (1971), Rodrigues (1984/1985), Rodrigues and Cabral (2002), Mello's (2002) revision of Rodrigues (1984/1985), and Walker et al.'s (2012) distance-based computational classification of the family. We find that the phylogenetic classification presented in this paper recovers many subgroups identified in previous classifications, and suggest that the divergences in higher-level structure may be in part explicable as differences emerging from the use of distance-based versus innovation-based subgrouping criteria.

Before moving to the comparisons, we introduce two terms originating in evolutionary biology that are useful for comparing evolutionary trees: 'monophyletic' and 'paraphyletic' groups. Monophyletic groups of languages are ones which contain all the (attested) descendants of a common ancestor and only descendants of that common ancestor;[19] monophyletic groups are thus identical to 'subgroups' in standard historical linguistic terminology. Critically, monophyletic groups are defined on the basis of shared innovations. Paraphyletic groups, or paraphyletic grades, are groups that contain only a subset of (attested) descendants of a common ancestor.[20] Paraphyletic groups are usually defined by shared retentions.

[19] For example, take Germanic, which contains all the attested daughters, and only the attested daughters of proto-Germanic

[20] For example, in standard reconstructions of Proto-Indo-European, Celtic and Italic together constitute a paraphyletic group, since there is no ancestral language of which they are the only the only descendants

We begin with Lemle (1971), whose classification of Tupí-Guaraní was based on the identification of several sound changes and the reconstruction of 220 words for 10 TG languages. Our analysis coincides with Lemle's in recovering the higher-order subgroup that includes the Southern and the Tupinambá subgroups, but we fail to recover Lemle's other subgroups.

We next turn to Rodrigues and Cabral (2002), which updates Rodrigues (1984/1985), and in many respects represents the state of the art of the traditional classifications of the family. It is therefore encouraging that our analysis exhibits considerable agreement with respect to the lower level groups, recovering five of the eight groups proposed by Rodrigues (1984/1985) and Rodrigues and Cabral (2002) (see Figure 2): Groups I, III, V, VI, and VII. Groups IV and VIII, however, are not recovered in our analysis in any coherent form, and Group II emerges as paraphyletic in our analysis. The way in which the traditional subgroups map onto our phylogenetic classification can be seen in Figure 2, where we have colored the languages belonging to each subgroup as follows: Group I in dark blue, Group II in orange, Group III in dark green, Group IV in purple, Group V in olive green, Group VI in light blue, Group VII in brown, and Group VIII in red.

Although there is considerable overlap in the lower-level groups between our and Rodrigues and Cabral's (2002) classifications, there is considerable divergence in the higher-level structure posited by the two proposals. This can be appreciated in Figure 4, which compares how the three first-order subgroups proposed by Rodrigues and Cabral (2002), color-coded red, green, and blue, map onto our classification. As we see in that figure, only one of their first-order subgroups, Group I – which corresponds to our Guaranian subgroup – is recovered as monophyletic in our analysis. Rodrigues and Cabral's (2002) two other first-order subgroups emerge in our analysis not as monophyletic groups, but as successive paraphyletic grades at the base of the Guaranian subgroup. Note also that although Group I emerges as monophyletic in our analysis, it is very deeply embedded in the tree, and not a first-order group, as in Rodrigues and Cabral (2002).

It is worth noting that incorrect inference of paraphyletic grades as monophyletic groups is a known weakness of distance-based methods (i.e., methods that group languages based on overall similarity), whether they are being used informally (e.g., through human "eyeballing"), or as computationally implemented algorithms. This weakness stems from the

fact that, as discussed in §4, distance-based methods include shared retentions as the basis for subgrouping. The differences in higher-level structure between Rodrigues and Cabral's (2002) classification and the one presented in this paper may thus result from the difference between subgrouping criteria: overall similarity in Rodrigues and Cabral's case, and innovations in ours. In fact, Rodrigues and Cabral's discussion of their classification makes it clear that they employed both shared innovations and retentions of phonological and morphological characters, suggesting that a possible fruitful direction for future research would be to take the phonological and morphological characters used by Rodrigues and Cabral (2002) and re-evaluate and re-optimize them on a tree. This may show that the evidence from these characters does not in fact contradict the lexical characters that we used in our analysis.

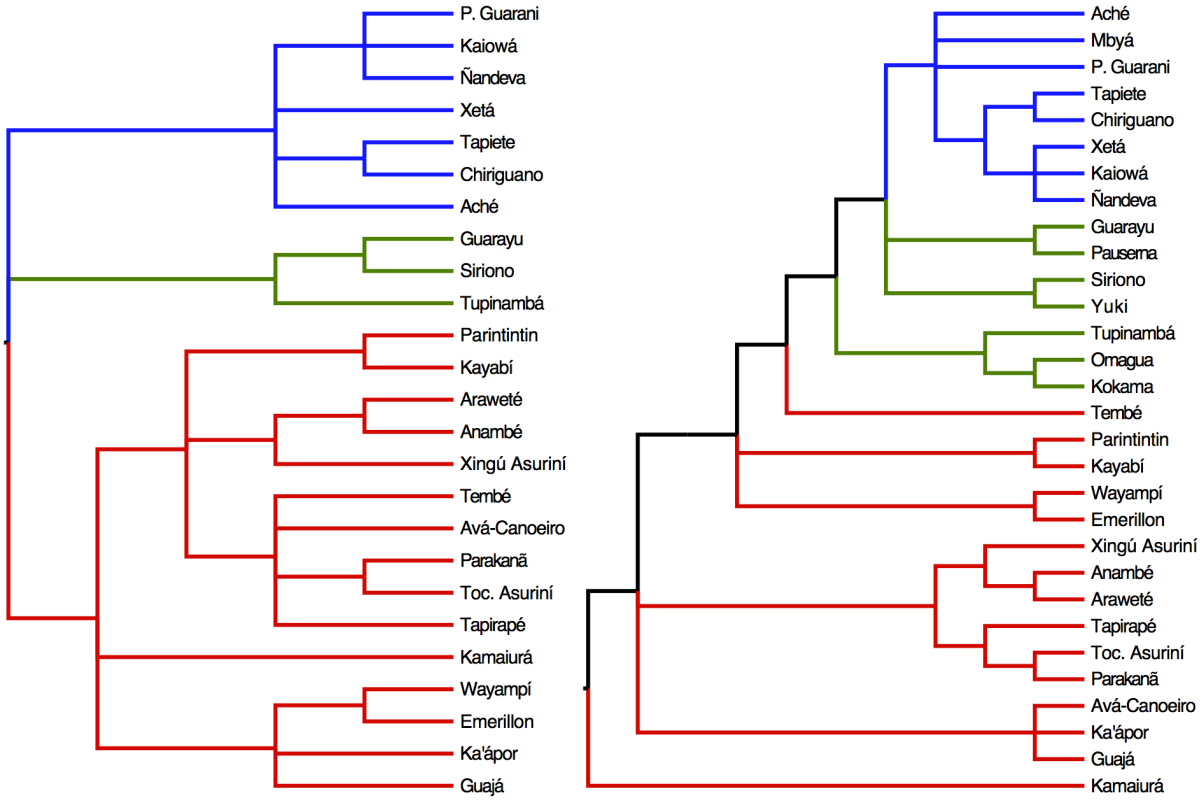

Figure 4: Higher Structure of Rodrigues and Cabral (2002) (left) and our classification (right) (corresponding areas of the trees are colored identically).

Mello (2002) reorganized Rodrigues' (1984/1985) eight subgroups into nine, splitting some and changing the subgroup membership of languages such as Kamaiurá, Parintintin, Guajá, and Xingú Asuriní. None of these changes are supported in our analysis.

Finally, we turn to Walker et al. (2012), who present a Neighbor-Joining tree of the entire Tupian stock based on a 40-item wordlist. Tupí-Guaraní is recovered as monophyletic in their analysis, but the internal structure of the family is strikingly different from both the results presented in this paper and previous classifications, with the exception of some of the low-level subgroups that all classifications have in common. Given the extremely small size of their dataset (less than a tenth the size of the one employed in this study), and the use of unreliable distance-based methods, the stark divergence of their results from both traditional classifications and our own phylogenetic one is not entirely surprising.

6 Conclusion

This study represents one of the largest efforts to date to clarify the relationships of Tupí-Guaraní languages both in terms of the number of languages included, as well as the dataset used. It also represents the first attempt to apply character-based phylogenetic methods to the study of Tupí-Guaraní. Based on a dataset of 543 lexical meanings, we propose a new internal classification of Tupí-Guaraní, which, although broadly compatible at lower-level subgroups with previous classifications, differs significantly in the higher-level topology. One of the most important differences of our results is that the widely recognized Southern subgroup is not a first-order subgroup as in previous classifications, but a deeply nested group. Also, other previously suggested higher-level groups are paraphyletic grades in our analysis. The position of the highly dispersed languages, deeply nested within the Tupí-Guaraní phylogeny, suggests an Amazonian origin for the Tupí-Guaraní languages.

A Ancestral State Reconstruction

Below are the cognate sets that are reconstructed as lost or gained for selected nodes on the majority-rule consensus tree of our analysis.[21] The list is not exhaustive for each node, but it includes all cognates that can be reconstructed as lost or gained on that node with high

[21] The cognate set names (e.g., dry6) are labels of convenience for cognate sets, and may include cognates that have experienced semantic shift away from the "common" meaning of the cognate set. For example, the Tupinambá exemplification for the dry6 set, *tuβir* means 'dust'. Note that the numerals following the common meaning of the cognate set serve to distinguish cognate sets with the same common meaning.

likelihood ($> 90\%$). Also, note that an inferred gain can be reversed in some of the daughter languages due to cognate loss, while an inferred loss may have also happened independently in more than one subgroups. The forms in the tables below come from different languages and are given as examples of a given cognate set; their orthographic representations have been standardized to the IPA.

Table 2: Ancestral State Reconstruction: Guaranian Subgroup

Cognate Set	Rec.	Example	Cognate Set	Rec.	Example
tapir1	gain	*mboreβi* (PG)	chief8	loss	*morerekwar* (tpn)
anteater6	gain	*kagware* (mby)	dark2	loss	*pihun* (tmb)
dry2	gain	*ɨpi* (PG)	dry6	loss	*tuβɨr* (tpn)
open1	gain	*ojei* (PG)	clean2	loss	*kɨtiŋok* (tpn)
deceive1	gain	*japu* (PG)			

Table 3: Ancestral State Reconstruction: Southern Subgroup

Cognate Set	Rec.	Example Form	Cognate Set	Rec.	Example Form
pineapple3	gain	*karagwata* (PG)	deer1	loss	*ɨtiː* (maw)
bat1	gain	*mbopi* (PG)	yellow5	loss	*tawa* (tpn)
digging stick1	gain	*sɨpe* (PG)	weak6	loss	*membek* (tpn)
follow1	gain	*moɲa* (PG)	far5	loss	*amõ* (tpn)
embrace3	gain	*kwãwa* (chi)	stop6	loss	*pɨk* (tpn)
howler monkey1	gain	*karaja* (PG)	throw4	loss	*ejtɨk* (tpn)
gourd3	loss	*kuj* (tpn)	light(v)3	loss	*mondɨk* (tpn)
lard4	loss	*kaβ* (tpn)	finish6	loss	*sɨk* (tpn)
howler monkey4	loss	*akɨkɨ* (tpn)			

Table 4: Ancestral State Reconstruction: Diasporic Subgroup

Cognate Set	Rec.	Example Form	Cognate Set	Rec.	Example Form
flute1	gain	*mimbɨ* (tpn)	pass2	gain	*pwan* (tpn)
bent, twisted3	gain	*βaŋ* (tpn)	be stinky1	gain	*timbor* (tpn)
far1	gain	*mombɨrɨ* (PG)	touch3	gain	*atõj* (tpn)
flow2	gain	*sururu* (chi)	shake2	gain	*mɨj* (tpn)
mourn1	gain	*apirõ* (tpn)			

Table 5: Ancestral State Reconstruction: Nuclear TG Subgroup

Cognate Set	Rec.	Example Form	Cognate Set	Rec.	Example Form
cheek3	gain	*atɨpɨ* (tpn)	full5	gain	*inɨsem* (tpn) (*sem*)
peanut1	gain	*manduβi* (tpn)	bury1	gain	*atiβ* (tpn)
old woman1	gain	*waiw* (kay)	breathe1	gain	*pɨtuʔẽ* (tpn)
spirit2	gain	*aɲaŋ* (tpn)	say1	gain	*mombeʔu* (tpn)
island1	gain	*ɨpaʔũ* (tpn)	sew1	gain	*mombɨβik* (tpn)
thing2	gain	*marã* (tpn)	stomach8	loss	*tɨʔa* (awe)
tasty1	gain	*ʔe* (tpn)	island4	loss	*ʔapem* (kam)
full4	gain	*inɨsem* (tpn) (compound)	pot10	loss	*jaʔapehẽ* (kam)

B Lexical Sources

- Aché: Heckart and Hill (2007); Hill (1983); Röβler (2008); Rößler (p.c.)
- Anambé: Silva Julião (2005)
- Araweté: Solano (2009); Viveiros de Castro (1992)
- Avá-Canoeiro: Borges (2006, 2007)
- Awetí: Corrêa da Silva (2010); Drude (2006, 2008, 2011); Drude (p.c.)
- Chiriguano: Dietrich (2007)
- Emerillon: Couchili et al. (2001); Gordon and Rose (2006); Queixalós (2001); Rose (2002, 2003, 2008, 2009, 2011); Rose (p.c.)
- Guajá: Cunha (1987); Magalhães (2006, 2007); Nascimento (2008)
- Guarayu: Armoye Urarepia (2009); Höller (1932)
- Ka'apor: Caldas (2009); Kakumasu and Kakumasu (1988); Lopes (2009)
- Kaiowá: Bridgeman (1961); Cardoso (2008); Harrison and Taylor (1971); Taylor (1984a,b)
- Kamaiurá: Drude (2011); Seki (1982, 1983, 1987, 1990, 2000a,b, 2007, 2010)
- Kayabí: Borges e Souza (2004); Dobson (1973, 1988, 1997)
- Kokama: Espinosa Pérez (1989); Faust (1959, 1972); Vallejos (2010); Vallejos (p.c.)
- Mawé: Corrêa da Silva (2010); Drude (2006); Franceschini (1999); Meira (p.c.)

- Mbyá: Dooley (2006)
- Ñandeva: Costa (2002, 2007); Dooley (1991)
- Omagua: O'Hagan (p.c.)
- Parakanã: da Silva (2003)
- Paraguayan Guaraní: Guasch (2003)
- Parintintin: Betts (1981); Pease (1968); Sampaio (1997)
- Pauserna: von Horn Fitz Gibbon (1955); Riester (1972)
- Siriono: Priest and Priest (1985)
- Tapiete: González (2005, 2008)
- Tapirapé: Almeida et al. (1983); Praça (2007)
- Tembé: Boudin (1978); Meira (p.c.)
- Tocantins Asuriní: Cabral and Rodrigues (2003); Harrison (1963, 1975); Nicholson (1978, 1982)
- Tupinambá: Lemos Barbosa (1951, 1970)
- Wayampí: Grenand (1989); Olson (1978)
- Xetá: Vasconcelos (2008)
- Xingú Asuriní: Nicholson (1978, 1982); Pereira (2009)
- Yuki: Garland (1978); Villafañe (2004)

References

Alekseyenko, Alexander V.; Christopher J. Lee; and Marc A. Suchard. 2008. Wagner and Dollo: A Stochastic Duet by Composing Two Parsimonious Solos. *Systematic Biology* 57(5):772–784.

Almeida, Antonio; Irmãzinhas de Jesus; and Luiz Gouvea de Paula. 1983. *A língua tapirapé*. Rio de Janeiro: Biblioteca Repográfica Xerox.

Armoye Urarepia, Celso. 2009. *Análisis de la lengua guarayo (tesina)*. Santa Cruz de la Sierra: ms.

Baele, Guy; Philippe Lemey; Trevor Bedford; Andrew Rambaut; Marc A. Suchard; and Alexander V. Alekseyenko. 2012. Improving the Accuracy of Demographic and Molecular Clock Model Comparison While Accommodating Phylogenetic Uncertainty. *Molecular Biology and Evolution* 29(9):2157–2167.

Balée, William. 1994. *Footprints of the Forest: Ka'apor Ethnobotany – The Historical Ecology of Plant Utilization by an Amazonian People*. New York: Columbia University Press.

Betts, La Vera. 1981. *Dicionário parintintin-português português-parintintin*. Cuiabá: Summer Institute of Linguistics (SIL).

Borges, Monica Veloso. 2006. *Aspectos fonológicos e morfossintáticos da língua avá-canoeiro (tupí-guaraní)*. PhD dissertation, Universidade Estadual de Campinas.

Borges, Monica Veloso. 2007. Posposições da língua avá-canoeiro (tupí-guaraní). *Línguas e Culturas Tupí*, edited by Ana Suelly Arruda Câmara Cabral and Aryon Dall'Igna Rodrigues, Campinas: Editora Curt Nimuendajú, 385–389.

Borges e Souza, Patrícia de Oliveira. 2004. *Estudos de aspectos da língua kaiabi (tupi)*. MA thesis, Universidade Estadual de Campinas, Campinas.

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.

Boudin, Max H. 1978. *Dicionário de tupi moderno: Dialeto tembé-ténetéhar do alto do rio Gurupi*. São Paulo: Conselho Estadual de Artes e Ciencias Humanas.

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

Bridgeman, Loraine I. 1961. Kaiwa (Guarani) Phonology. *International Journal of American Linguistics* 27(4):329–334.

Cabral, Ana Suelly Arruda Câmara. 1995. *Contact-Induced Language Change in the Western Amazon: The Non-Genetic Origin of the Kokama Language*. PhD dissertation, University of Pittsburgh.

Cabral, Ana Suelly Arruda Câmara and Aryon Dall'Igna Rodrigues. 2003. *Dicionário asuriní do tocantins-português*. Belém: Universidade Federal do Pará.

Caldas, Raimunda Benedita Cristina. 2009. *Uma proposta de dicionário para a língua ka'apór*. PhD dissertation, Universidade de Brasília.

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

Cardoso, Valéria Faria. 2008. *Aspectos morfossintáticos da língua kaiowá (guaraní)*. PhD dissertation, Universidade Estadual de Campinas.

Chang, Will; Chundra Cathcart; David Hall; and Andrew Garrett. 2015. Ancestry-constrained Phylogenetic Analysis Supports the Indo-European Steppe Hypothesis. *Language* 91(1):194–244.

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.

Costa, Consuelo De Paiva Godinho. 2002. A nasalização em nhandewa-guarani. *Línguas Indígenas Brasileiras: Fonologia, Gramática e História*, vol. 1, edited by Ana Suelly Arruda Câmara Cabral and Aryon Dall'Igna Rodrigues, Belém: Editora Universitária, Universidade Federal do Pará, 403–412.

Costa, Consuelo De Paiva Godinho. 2007. *Apyngwa rupigwa: Nasalização em nhandewa-guaraní*. PhD dissertation, Universidade Estadual de Campinas.

Couchili, T.; D. Maurel; and Francisco Queixalós. 2001. Classes de lexèmes en émérillon. *Amerindia* 26/27:173–208.

Crowley, Terry and Claire Bowern. 2010. *An Introduction to Historical Linguistics*. Oxford: Oxford University Press.

Cunha, Péricles. 1987. *Análise fonêmica preliminar da língua guajá*. MA thesis, Universidade Estadual de Campinas.

da Silva, Gino Ferreira. 2003. *Construindo um dicionário parakanã-português*. MA thesis, Universidade Federal do Pará.

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

Dietrich, Wolf. 2007. Nuevos aspectos de la posición del conjunto chiriguano (guaraní del chaco boliviano) dentro de las lenguas tupí-guaraníes bolivianos. *Lenguas indígenas de América del Sur: estudios descriptivo-tipológicos y sus contribuciones para la lingüística teórica*, edited by Andrés Romero-Figueroa; Ana Fernández Garay; and Ángel Corbera Mori, Caracas: Universidad Católica Andrés Bello, 9–18.

Dobson, Rose M. 1973. Notas sobre substantivos do kayabí. *Serie Lingüística* 1:30–56.

Dobson, Rose M. 1988. *Aspectos da língua kayabí*. No. 12 in Serie Lingüística, Cuiabá: Summer Institute of Linguistics (SIL).

Dobson, Rose M. 1997. *Gramática prática com exercícios da língua kayabí*. Cuiabá: Summer Institute of Linguistics (SIL).

Dooley, Robert A. 1991. *Apontamentos preliminares sobre ñandéva guaraní contemporâneo*. ms.

Dooley, Robert A. 2006. *Léxico guaraní, dialeto mbyá com informações úteis para o ensino médio, a aprendizagem e a pesquisa lingüística*. Cuiabá: Summer Institute of Linguistics (SIL).

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. 2008. Nasal Harmony in Awetí and the Mawetí-Guaraní Family (Tupí). *Amerindia* 32:239–267.

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.

Drummond, Alexei J. and Andrew Rambaut. 2007. BEAST: Bayesian Evolutionary Analysis by Sampling Trees. *BMC Evolution Biology* 7(1):214.

Drummond, Alexei J.; Marc A. Suchard; Dong Xie; and Andrew Rambaut. 2012. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* 29(8):1969–1973.

Espinosa Pérez, Lucas. 1989. *Breve diccionario analítico castellano-tupí del Perú: Sección cocama*. Iquitos: Instituto de Investigaciones de la Amazonía Peruana (IIAP); Centro de Estudios Teológicos de la Amazonía (CETA).

Faust, Norma. 1959. Vocabulario breve del idioma cocama (tupí). *Perú Indígena* 8(18-19):150–158.

Faust, Norma. 1972. *Gramática cocama: Lecciones para el aprendizaje del idioma cocama*. No. 6 in Serie Lingüística Peruana, Lima: Summer Institute of Linguistics (SIL).

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*, 9079–9084.

Franceschini, Dulce. 1999. *La langue sateré-mawé: Description et analyse morphosyntaxique*. PhD dissertation, Université Paris VII.

Garland, Mary. 1978. *Diccionario yuki-inglés*. ms.

González, Hebe Alicia. 2005. *A Grammar of Tapiete (Tupí-Guaraní)*. PhD dissertation, University of Pittsburgh.

González, Hebe Alicia. 2008. Una aproximación a la fonología del tapiete (tupí-guaraní). *LIAMES* 8:7–43.

Gordon, Matthew and Françoise Rose. 2006. Émérillon Stress: A Phonetic and Phonological Study. *Anthropological Linguistics* 48(2):132–168.

Gray, Russell D. and Quentin D. Atkinson. 2003. Language-tree divergence times 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.

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.

Grenand, Françoise. 1989. *Dictionnaire wayãpi/français*. Paris: Peeters/Selaf.

Guasch, Antonio. 2003. *Diccionario básico guaraní-castellano, castellano guaraní*. Asunción: CEPAG.

Harrison, Carl H. 1963. *Pedagogical Information and Drills for the Asuriní Language*. ms.

Harrison, Carl H. 1975. Gramática asuriní: Aspectos de uma gramática transformacional e discursos monologados da língua asuriní, família tupí guaraní. *Serie Lingüística* 4.

Harrison, Carl H. and John M. Taylor. 1971. Nasalization in Kaiwá. *Tupí Studies I*, edited by David Bendor-Samuel, Summer Institute of Linguistics (SIL), 15–20.

Heckart, Claudia and Kim Hill. 2007. *Aché*. Intercontinental Dictionary Series.

Hill, Kim. 1983. Neotropical Hunting Among the Aché of Eastern Paraguay. *Adaptive Responses of Native Amazonians*, edited by Raymond B. Hames and William T. Vickers, New York: Academic Press, 139–188.

Höller, Alfredo. 1932. *Guarayo-Deutsches Wörterbuch*. Guarayos: Verlag der Missionsprokura der P.P. Franziskaner, Hall in Tirol.

Huelsenbeck, John P. and Fredrik Ronquist. 2001. MRBAYES: Bayesian Inference of Phylogenetic Trees. *Bioinformatics* 17(8):754–755.

Jensen, Cheryl. 1989. *O desenvolvimento histórico da língua wayampí*. Campinas: Editora da UNICAMP.

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.

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

Kakumasu, James Y. and Kiyoko Kakumasu. 1988. *Dicionário por tópicos kaapor-português*. Brasília: Summer Institute of Linguistics (SIL).

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

Kass, Robert E. and Adrian E. Raftery. 1995. Bayes Factors. *Journal of the American Statistical Association* 90(430):773–795.

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.

Lemle, Miriam. 1971. Internal Classification of the Tupí-Guaraní Linguistic Family. *Tupí Studies I*, edited by David Bendor-Samuel, Summer Institute of Linguistics (SIL), 107–129.

Lemos Barbosa, Antônio. 1951. *Pequeno vocabulário tupí-português*. Rio de Janeiro: Livraria São José.

Lemos Barbosa, Antônio. 1970. *Pequeno vocabulário português-tupí*. Rio de Janeiro: Livraria São José.

Lopes, Mário Alexandre Garcia. 2009. *Aspectos gramaticais da língua ka'apor*. PhD dissertation, Universidade Federal e Minas Gerais.

Maddison, Wayne and David R. Maddison. 2007. *Mesquite: A Modular System for Evolutionary Analysis*. URL `http://mesquiteproject.org`.

Magalhães, Marina Maria Silva. 2006. Harmonia vocálica como processo desencadeador de mudanças estruturais na língua guajá. *Estudos da Língua(gem)* 4(2):67–75.

Magalhães, Marina Maria Silva. 2007. *Sobre a morfologia e a sintaxe da língua guajá (família tupí-guaraní)*. PhD dissertation, Universidade de Brasília.

Mello, Antônio Augusto Souza. 2000. *Estudo histórico da família lingüística tupí-guaraní: Aspectos fonológicos e lexicais*. PhD dissertation, Universidade Federal de Santa Catarina.

Mello, Antônio Augusto Souza. 2002. Evidencias fonológicas e lexicais para o subagrupamento interno tupí-guaraní. *Línguas Indígenas Brasileiras: Fonologia, Gramática*

e História, vol. 1, edited by Ana Suelly Arruda Câmara Cabral and Aryon Dall'Igna Rodrigues, Belém: Editora Universitária, Universidade Federal do Pará, 338–342.

Michael, Lev. 2014. On the Pre-Columbian Origin of Proto-Omagua-Kokama. *Journal of Language Contact* 7(2):309–344.

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.

Nascimento, Ana Paula Lion Mamede. 2008. *Estudo fonético e fonológico da língua guajá*. MA thesis, Universidade de Brasília.

Nicholson, Velda C. 1978. *Aspectos da língua assurini*. ms.

Nicholson, Velda C. 1982. Breve estudo da língua do xingú. *Ensaios Lingüísticos* 5.

Noelli, Francisco S. 2008. The Tupi Expansion. *The Handbook of South American Archeology*, edited by Helaine Silverman and William H. Isbell, New York: Springer, 659–670.

Olson, Gary Paul. 1978. Descrição preliminar de orações wayãpi. *Ensaios Lingüísticos* 3.

Pease, Helen. 1968. *Parintintin Grammar*. Porto Velho: Associação Internacional de Lingüística.

Pereira, Antônia Alves. 2009. *Estudo morfossintático do asuriní do xingú*. PhD dissertation, Universidade Estadual de Campinas.

Praça, Walkíria Neiva. 2007. *Morfossintaxe da língua tapirapé*. PhD dissertation, Universidade de Brasília.

Priest, Perry N. and Anne M. Priest. 1985. *Diccionario siriono y castellano*. Cochabamba: Summer Institute of Linguistics (SIL).

Queixalós, Francisco. 2001. Le suffixe référentiant en émérillon. *Des noms et des verbes en tupi-guarani: État de la question*, edited by Francisco Queixalós, Munich: LINCOLM Europa, 115–132.

Rambaut, Andrew and Alexei J. Drummond. 2007. *Tracer v1.4*. URL `http://beast.bio.ed.ac.uk/Tracer`.

Riester, Jürgen. 1972. *Die Pauserna-Guarasu'wä: Monographie eines Tupí-Guaraní-Volkes in Ostbolivien*. St. Augustin bei Bonn: Verlag des Anthropos-Instituts.

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. 1978. A língua dos índios xetá como dialeto guaraní. *Cadernos de Estudos Lingüísticos* 1:7–11.

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

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.

Röβler, Eva-Maria. 2008. *Aspectos da gramática achê: Descrição e reflexão sobre uma hipótese de contato*. MA thesis, Universidade Estadual de Campinas.

Ronquist, Fredrik and John P. Huelsenbeck. 2003. MrBayes 3: Bayesian Phylogenetic Inference Under Mixed Models. *Bioinformatics* 19(12):1572–1574.

Rose, Françoise. 2002. Le problème de la nasalité dans l'inventaire phonologique de l'émérillon. *Amerindia* 26/27:147–172.

Rose, Françoise. 2003. Le marquage des personnes en emérillon (tupí-guaraní): Un système d'accord hiérarchique. *Faits de Langues* 21(2):107–120.

Rose, Françoise. 2008. A Typological Overview of Emerillon, a Tupí-Guaraní Language from French Guiana. *Linguistic Typology* 12:431–460.

Rose, Françoise. 2009. A Hierarchical Indexation System: The Example of Emerillon (Teko). *New Challenges in Typology: Transcending the Borders and Refining the Distinctions*, edited by Patricia Epps and A. Arkhipov, Berlin: Mouton de Gruyter, 63–84.

Rose, Françoise. 2011. *Grammaire de l'émérillon teko: Une langue tupi-guarani de guyane française*. Louvain-Paris: Peeters.

Sampaio, Wany Bernardete de Araujo. 1997. *Estudo comparativo sincrônico entre o parintintin (tenharim) e o uru-eu-uau-uau (amondava): Contribuições para uma revisão na classificação das línguas tupí-kawahib*. MA thesis, Universidade Estadual de Campinas.

Schleicher, Charles Owen. 1998. *Comparative and Internal Reconstruction of Proto-Tupí-Guaraní*. PhD dissertation, University of Wisconsin, Madison.

Seki, Lucy. 1982. Marcadores de pessoa no verbo kamaiurá. *Cadernos de Estudos Lingüísticos* 3:22–40.

Seki, Lucy. 1983. Observações sobre variacão sociolingüística kamayurá. *Cadernos de Estudos Lingüísticos* 4:73–87.

Seki, Lucy. 1987. Para uma caracterização tipológica do kamaiurá. *Cadernos de Estudos Lingüísticos* 12:15–24.

Seki, Lucy. 1990. Kamaiurá (Tupí-Guaraní) as an Active-Stative Language. *Amazonian Linguistics: Studies in Lowland South American Languages*, edited by Doris L. Payne, Austin: University of Texas Press, 367–391.

Seki, Lucy. 2000a. Aspectos diacrônicos da língua kamaiurá (tupí-guaraní). *Linguistica Romanica et Indiana: Festschrift für Wolf Dietrich zum 60. Geburtstag*, edited by Bruno Staib, Tübingen: Gunter Narr Verlag, 565–581.

Seki, Lucy. 2000b. *Gramática do kamaiurá: Língua tupí-guaraní do alto Xingu*. Campinas: Editora da UNICAMP.

Seki, Lucy. 2007. Partículas e tipos de discurso em kamaiurá (tupí-guaraní). *Lenguas Indígenas de América del Sur: Estudios Descriptivo-tipológicos y sus Contribuciones para la Lingüística Teórica*, edited by Andrés Romero-Figueroa; Ana Fernández Garay; and Ángel Corbera Mori, Caracas: Universidad Católica Andrés Bello, 145–157.

Seki, Lucy. 2010. *Jene ramỹjwena juru pytsaret: O que habitava a boca de nossos ancestrais*. Rio de Janeiro: Museu do Índio - FUNAI.

Silva Julião, Maria Risolêta. 2005. *Aspects morphosyntaxiques de l'anambe*. PhD dissertation, Université de Toulouse, Le Mirail.

Soares, Marília Facó and Yonne Leite. 1991. Vowel Shift in the Tupí-Guaraní Language Family: A Typological Approach. *Language Change in South American Indian Languages*, edited by Mary Ritchie Key, Philadelphia: University of Pennsylvania Press, 36–53.

Solano, Eliete De Jesus Bararuá. 2009. *Descrição gramatical da língua araweté*. PhD dissertation, Universidade de Brasília.

Taylor, John M. 1984a. A interrogação na língua kaiwá. *Estudos sobre Línguas Tupí do Brasil*, edited by Robert A. Dooley, Brasília: Summer Institute of Linguistics (SIL).

Taylor, John M. 1984b. Marcação temporal na língua kaiwá. *Estudos sobre Línguas Tupí do Brasil*, edited by Robert A. Dooley, Brasília: Summer Institute of Linguistics (SIL).

Vallejos, Rosa. 2010. *A Grammar of Kokama-Kokamilla*. PhD dissertation, University of Oregon.

Vasconcelos, Eduardo Alves. 2008. *Aspectos fonológicos da língua xetá*. MA thesis, Universidade de Brasília.

Villafañe, Lucrecia. 2004. *Gramática yuki: Lengua tupí-guaraní de Bolivia*. Tucumán: Ediciones del Rectorado, Universidad Nacional de Tucumán.

Viveiros de Castro, Eduardo. 1992. *From the Enemy's Point of View: Humanity and Divinity in an Amazonian Society*. Chicago: University of Chicago Press.

von Horn Fitz Gibbon, Friedrich. 1955. *Breves notas sobre la lengua de los indios pausernas: El üaradu-ñé-e (un dialecto tupí-guaraní en el oriente de Bolivia)*. Publicaciones de la Sociedad de Estudios Geográficos e Históricos, Santa Cruz de la Sierra: Imprenta Emilia.

Walker, Robert S. and Lincoln A. Ribeiro. 2011. Bayesian Phylogeography of the Arawak Expansion in Lowland South America. *Proceedings of the Royal Society* 278(1718):2562–2567.

Walker, Robert S.; Søren Wichmann; Thomas Mailund; and Curtis J. Atkisson. 2012. Cultural Phylogenetics of the Tupi Language Family in Lowland South America. *PLoS ONE* 7(4):e35,025.

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*.

Wiens, John J. 2003. Missing Data, Incomplete Taxa, and Phylogenetic Accuracy. *Systematic Biology* 52(4):528–538.

Wiens, John J. and Matthew C. Morrill. 2011. Missing Data in Phylogenetic Analysis: Reconciling Results from Simulations and Empirical Data. *Systematic Biology* 60(5):719–731.