Subgrouping in the Tupí-Guaraní Family: A Phylogenetic Approach*

NATALIA CHOUSOU-POLYDOURI and VIVIAN WAUTERS
University of California, Berkeley

1 Introduction

In recent years, interdisciplinary collaboration between linguists and biologists has led to new insights in comparative linguistics. By analyzing linguistic data using phylogenetic methods of species subgrouping, we can quantify statistic probabilities of various subgrouping hypotheses and enrich our understanding of the diachronic relationships between descendant languages in a family. Even as the preliminary results of this work have developed important strands of inquiry (Atkinson and Gray 2003; Rexová et al. 2006; Dunn 2009), many have criticized the method, indicating essentially that "we still do not have evidence that any of these methods [of phylogenetic analysis] is capable of accurate estimation of linguistic phylogenies" (Nichols and Warnow 2008:814; see also Heggarty 2006).

This paper presents a preliminary analysis of the Tupí-Guaraní language family using phylogenetic methods. Through use of parsimony and Bayesian analyses on a set of lexical items divided into cognates, we conclude that phylogenetic methods do produce useful and interesting results that will inform more traditional reconstruction. We also discuss the many potential confounds that are present in the current analysis and their solutions, and point to future research directions that will make further use of this data in conjunction with a "by hand" reconstruction of the Tupí-Guaraní proto-language.

1.1 Language Background

The Tupí-Guaraní language family forms one major subgroup of the Tupian stock, which is one of the largest macro families in South America (Derbyshire 1994). By the time of European invasion, the Tupian languages were spread throughout the Amazon and beyond, from the Andes to the Atlantic Ocean, mostly following the complex river systems of the regions (Lathrap 1970). Following contact, many languages were lost, and many others moved location, either through forced re-location or in the process of fleeing the settlers (for a discussion of these processes on Omagua, see Michael 2010). However, the recordable remnants of this widely spread family are visible in the dispersion of Tupí-Guaraní languages, as seen in Figure 1.

1.2 Previous Study

While South America is one of the linguistically least understood areas of the world (Dixon and Aikhenvald 1999), the Tupí stock, and particularly the Tupí-Guaraní family, has a somewhat es-

*Many thanks to Lev Michael, who runs the Tupí-Guaraní Comparative Project, as well as colleagues Keith Bartolomei, Zachary O'Hagan, and Michael Roberts, each of whom is responsible for collecting all the lexical items for a portion of the languages.

Figure 1: Tupí-Guaraní language map

tablished tradition of description ascribed to it (Soares and Leite 1991:36). In addition to the description efforts on various daughter languages of the family, Rodrigues (1958, 1984, 2007) and Rodrigues and Cabral (2002) all focus on various aspects of reconstructing proto-Tupí-Guaraní. Broader morphosyntactic reconstruction efforts appear to be based on Lemle (1971). This reconstruction is referenced and utilized in much of Rodrigues' work, and contributed to the influential set of reconstructions by Cheryl Jensen (1989, 1998). The latter of the two, Jensen (1998), titled "Comparative Tupí-Guaraní Morphosyntax," is the source of most linguists' understanding of subgrouping in the Tupí-Guaraní family. Both Jensen (1998) and Rodrigues (1984) posit only a single level of subgroupings of the Tupí-Guaraní languages; Rodrigues' grouping, copied into Jensen (1998) is shown in Table 1, with the languages considered in this work in bold.[1] While this is not the only subgrouping that has been proposed, the more recent Rodrigues and Cabral (2002), which included more detailed subgroups within the eight established groups, differed only minimally from its antecendent: Kayabí is considered part of group VI and not V. This is also the case in Etnolingüística (2011).

GROUP	LANGUAGE
I	Old Guaraní; Mbyá Guaraní [Mbyá]; **Xetá**; Ñandeva; Kaiwá [Kaiowá]; **Paraguayan Guaraní**; Guayakí; **Tapieté**; Chiriguano; Izoceño
II	**Guarayu**; Sirionó; Hora; **[Yuki]**
III	**Tupinambá**; Língua Geral Paulista; Língua Geral Amazônica (Nheengatu); **Cocama [Kokama]**, **Cocamilla [Kokamilla]**; **Omagua**
IV	**Tocantins (or Trocará) Assuriní [Asuriní do Tocantins]**; **Tapirapé**; **Ava (Canoeiro)**; Tocantins Suruí (Akewere); **Parakanã**; **Guajajára** ; **Tembé [Tenehetara]**
V	**Kayabí**; **Xingu Assuriní [Asuriní do Xingu]**; Araweté
VI	**Parintintín**; Tupí-Kwahíb; Apiaká
VII	**Kamaiurá**
VIII	Takunyapé; Emerillon; Urubú-Kaapor; Wayampi; Amanayé; Anambé; Turiwara; **Guajá**

Table 1: Jensen (1998) Tupí-Guaraní language subgrouping

[1] Names in square brackets represent the names of the languages as used in this study, when different from those used in the source material. Yuki is not included in this subgrouping, but is classified in Etnolingüística (2011) as a member of Group II. Also, our analysis treats Guajajara and Tembé as dialects that can be subsumed under the title Tembé-Tenetehara, often shortened to Tembé. Similarly, Kokama and Kokamilla are dialects that will be referred to under the single name Kokama.

Despite these published reconstruction efforts, none of the previously mentioned sources have actually reconstructed proto-Tupí-Guaraní using rigorous historical linguistics methods like the Comparative Method. What seems more accurate is that Lemle (1971) reconstructed a phoneme system of proto-Tupí-Guaraní by deduction from segments in the daughter languages, and used that phoneme set to establish proto-forms of some 200 words. While this is only one in the list of reconstructions, it further appears that most subsequent work on the proto-language has used the Lemle (1971) data as its basis. While these extant reconstructions do seem (from observation) to be generally correct, the lack of precision inherent in the methods used makes current scholarship on proto-Tupí-Guaraní ultimately uninformative.

In contrast to these "reconstructions," there have been two published works on the diachronic relationship of languages in the Tupí-Guaraní family that make conclusions based on data rather than intuition. The first of these is "More Evidence for an Internal Classification of Tupí-Guaraní Languages" by Wolf Dietrich (1990). Unlike the other reconstructions, Dietrich does not attempt to make a tree structure for the diversification from proto-Tupí-Guaraní, but rather, using 29 languages, he examines phonological and morphological characteristics of the languages, and, using a summary of pairwise combinations between languages, posits a gradient analysis of each Tupí-Guaraní language from "conservative" to "innovative." Under this analysis, Dietrich does not propose any proto-forms but groups the languages based on their level of conservativeness into a number of low-level subgroups. These results differ from the other reconstructions in some important ways. Most notably, Kokama (Cocama) does not form a subgroup with Tupinambá. Also, the two Group 6 languages of previous reconstructions that were included in the sample do not pattern together, but rather Kayabí (Group 6) appears to be much closer to Tapirapé (a Group 4 language) and Kamaiurá (Group 7) than to Asuriní do Xingu (Group 6) (Dietrich 1990:116). Notably, Dietrich freely admits that while his study is able to indicate similarities between language pairs and small groups, it does not constitute a fully developed description of the subgrouping in Tupí-Guaraní, which he considered to be "far from clear" (Dietrich 1990:116).[2]

The second of these reconstructions is "A Historical Study of the Tupí-Guaraní Language Family" (Estudo histórico da família lingüística tupí-guaraní; Mello 2000). This study actually reconstructs a number of cognate sets based on regular sound correspondences, and forms subgrouping hypotheses based on sound changes as well as cognate isoglosses; the grouping is shown in Table 2.[3]

[2] There is another study that reconstructs proto-Tupí-Guaraní forms based on what appear to be more principled methodologies, which is Schleicher (1998). This reconstruction does not, however, propose any subgrouping hypothesis, instead citing Dietrich (1990) as the best attempt.

[3] Languages included in our current sample appear in bold and languages that will eventually be included appear in italics. For ease of comparison, the language names have been regularized from Mello's labels to those used throughout the rest of this study.

GROUP	SUBGROUP	LANGUAGE
I	A	*Mbyá*, **Paraguayan Guaraní**, *Old Guaraní*
	B	*Chiriguano*, Chané, Izoceño
	C	Guayaki
	D	**Xetá**
II		*Sirionó*
III		**Guarayú**
IV	A	**Parintintin**, Amundava, Urueuewauwau
	B	Tenharín, Karipúna
V		**Kayabí**, **Kamaiurá**, *Apiaká*
VI	A	**Asuriní do Tocantins**, **Parakanã**, Suruí
	B	**Tembé**
	C	**Tapirapé**
	D	**Asuriní do Xingu**
VII		**Guajá**, *Araweté*, Anambé, Aurê e Aura
VIII		*Wayampí do Amapari/Jarí*, *Emerillon*, *Urubu-Kaapór*
IX		**Tupinambá**, **Kokama**, Língua Geral Amazônica

Table 2: Mello (2000) Tupí-Guaraní language subgrouping

While Mello (2000) does not include a number of the languages in our current study in his subgrouping analysis, there are a number of significant and interesting differences between his analysis and those of the other subgroupings.[4] Most notably, Mello adds a Group 9, in which he puts Tupinambá and Kokama, which corresponds to Group 3 in the other analyses. He also subdivides the traditional Group 2 into two groups, one with Sirionó and the other with Guarayú. Furthermore, his Group 6 corresponds to an amalgamation of the other analyses' Groups 4 and 6, with the notable exception of Parintintin. Finally, he clusters Kayabí and Kamaiurá together unlike previous analyses. These significant differences, particularly in light of the potentially more reliable data collection and analysis methods used in Mello (2000) make it particularly important as a basis for comparison in our current study.

[4] We assume that Omagua would cluster with Kokama in Group IX, but Tapieté, Yuki, and Ava Canoeiro are also missing, with few clues as to their distribution.

2 Methodology

2.1 Data Collection and Diversity

Linguists traditionally use the Comparative Method, both for reconstruction of proto-segments and words, as well as to determine subgrouping within language families. Languages are grouped according to shared innovations, developments from the proto-language that appear in a subset of languages. Subgrouping using the Comparative Method is based on the assumption that it is more parsimonious to assume that languages that exhibit shared innovations inherited those innovations from an intermediate proto-language, rather than proposing that each language developed the innovation independently. Of course, this is complicated by possible contact effects and diffusion of features, but traditional reconstruction assumes that such effects are negligible compared to the main effects of shared innovations. In either case, shared retention is not indicative of any particular subgrouping.

While the Comparative Method tests theories about language relationships, phylogenetics assumes relatedness, and provides subgroupings using an optimality criterion. Thus, a requirement of using phylogenetic methods in linguistics is that one considers the languages to be related at some level. In the case of the Tupí-Guaraní family this is fairly unproblematic. Published work on the family is, for the most part, in agreement about which languages form part of the family. There are two notable exceptions. First, some authors, such as Schleicher (1998) place Sirionó outside of the Tupí-Guaraní family. More significantly, both Omagua and Kokama, two languages often treated as dialectal variants of one another, have previously been analyzed as sufficiently mixed in heritage to not qualify as strictly Tupí-Guaraní (Cabral 1995; Schleicher 1998; see also Dietrich 1990).

For the phylogenetic analysis, we used a dataset of 20 languages, 18 of which are Tupí-Guaraní and two of which are Tupian. Based on the extant subgrouping hypotheses, we made sure to include a language from each of the eight groups, with some groups represented by more than a single language. Within each group, representative languages were chosen to be included in the analysis based on the availability and quality of primary sources. Not all Tupí-Guaraní languages have sufficient description to gather the necessary data for our purposes, and we have not completed data collection for all languages for which it is possible.[5] The two Tupian languages were chosen to be outgroup languages because they are conventionally ascribed to sit just outside of the Tupí-Guaraní cluster in the larger Tupian family. Awetí is sister to Tupí-Guaraní and Satere-Mawé is sister to the Awetí-Tupí-Guaraní subgroup.

The reason for including data outside the Tupí-Guaraní family is so that we are able to root the tree. As discussed clearly in Dunn et al. (2008), rooting a tree consists of picking up an unrooted tree indicating relationships between languages and "suspend[ing] it at different points" (Dunn et al. 2008:723). By including two outgroup languages, we know from which point to suspend our tree: from the point at which the more distant outgroup language connects to the unrooted tree. The reason to include at least 2 outgroup languages is in order to test the monophyly of Tupí-Guaraní (i.e., if the Tupí-Guaraní languages form a subgroup, leaving the second outgroup

[5] This will naturally lead us to re-test the data using similar methods once our sample is complete with all 27 languages that do have sufficient data represented.

language outside).

For each of the 20 languages included in the study, we collected data using primary sources such as wordlists, dictionaries, grammars, and personal fieldnotes.[6] We chose which lexical items to include in our list by taking a 200-item Swadesh list which was expanded with words that are culturally important in the area (animal names, food items, tools, weapons, other artifacts, and verbs) and words that had reconstructed protoforms in the previous reconstruction work. Words in the Swadesh list that were irrelevant for climatic, cultural, or linguistic reasons were removed. We then searched the sources in each language for words bearing the meaning of our chosen lexical list items as well as words with closely related meanings. Our final dataset for this study included 572 word meanings (5 numbers, 52 body part terms, 25 food terms, 52 animal terms, 42 kinship terms, 7 color terms, 18 time-related terms, 41 natural environment terms, 49 artifacts, 63 adjectives, 13 non-adjectival descriptive words, and 205 verbs). The full list of meanings and the corresponding words found for each language are available upon request.

2.2 Character Coding

Characters used for computational phylogenetic methods need to follow a number of assumptions in order for the methods to be trustworthy. First of all, researchers need to be able to compare characters across languages, i.e., to formulate hypotheses of homology.[7] It must be noted here that character coding is a hypothesis of homology for many characters to be tested with an optimality criterion (parsimony, maximum likelihood). In the absence of evidence to the contrary, the researcher accepts common ancestry as the simplest explanation for the observed similarity. Characters must also be "heritable," which is not a trivial concept in linguistics. At a microscale, most linguists accept that language is not "inherited," but acquired in a more or less unique way by an individual based on his or her experience within a linguistic environment. At a more macroscopic level, though, the language of a speech community exhibits "heritability" as its characteristics are retained, modified, or lost. The second main assumption for characters used in a phylogenetic analysis is that they are independent, i.e., that they evolve separately from each other. Thus, their congruence is evidence for shared history.

There are two ways that lexical data can be coded for phylogenetic analyses: treating each etymon as a character with states present or absent in each language, or treating each meaning as a

[6] The sources consulted for the data from each language are as follows: Asuriní do Tocantins - Harrison (1975), Harrison (2009); Asuriní do Xingu - Nicholson (1982), Pereira (2009); Avá Canoero - Borges (2006), Borges (2007); Awetí - Borella (2000), Corrêa da Silva (2010), Drude (2011); Guajá - Cunha (1987), Magalhães (2006), Magalhães (2007), Nascimento (2008); Guarayú - Armoye (2009); Kamaiurá - Seki (1982), Seki (1983), Seki (1987), Seki (1990), Seki (2000), Seki (2007); Kayabí - Dobson (1988), Dobson (1997); Kokama - Faust (1959), Faust (1971), Espinosa (1935), Vallejos (2010); Omagua - (O'Hagan et al. 2011); Paraguayan Guaraní - Guasch (2003); Parakaná - da Silva (2003); Parintintin - Pease (1968), Sampaio (1977), Betts (1981); Sateré-Mawé - Franceschini (2000), Corrêa da Silva (2010); Tupinambá - Lemos Barbosa (1970); Tapieté - González (2005), González (2008); Tapirapé - Almeida et al. (1983), Praça (2007); Tembé-Tenetehara - Boudin (1978); Wayampí - Olson (1978); Yuki - Villafañe (2004); Xetá - Vasconcelos (2008).

[7] In biology, homologous morphological structures, genes, or nucleotides can be defined as being manifestations of the same morphological structure, gene, or nucleotide present in the common ancestor of the group studied (e.g., the limbs of amphibians, reptiles, birds, and mammals are homologous and they are modifications of the limbs of the common ancestor of all tetrapods).

character with states corresponding to cognate sets within this semantic slot.

The first coding method is based on etyma, which are unambiguously "inherited," but which might not be independent. The independence of etyma which are close to each other in semantic space is essentially a question on the prevalence or not of semantic shift: When a new etymon enters a semantic slot, will the old etymon disappear or change meaning? The answer is that both outcomes can happen. If semantic shift happens very rarely, then etyma are clearly non-independent and should not be used as the basis for characters. However, if semantic shift is more or less common, then etyma are independent to a certain extent and there is subgrouping information that can be lost if it is not taken into account.[8] This coding scheme results in binary characters with states (present/absent) that are comparable across many characters, making it easy to apply a common evolutionary model and use likelihood-based and Bayesian methods.

The second coding method is based on the homology of meanings, which are not as intuitively "inherited" as etyma, but they are independent. In fact, "meanings" in the form of body parts, animals, natural phenomena, and others exist outside language and language applies a word to them. The assumption here is that these "meanings" are comparable across languages.[9] This coding scheme results in multistate characters (each state corresponds to a cognate set) without the implication of a unified mechanism for state changes. This makes the characterization of models, necessary for likelihood-based and Bayesian methods, more complicated and could result in overparameterization.[10] Nevertheless, multistate characters of this type are ideal for parsimony methods.

Most previous works (Greenhill and Gray 2005; Gray et al. 2009; Atkinson and Gray 2003) have used essentially the second type of coding (but see Rexová et al. 2006), followed by a step of binary recoding, where each cognate set within a semantic slot is considered a character and is coded as present or absent.[11] This step introduces a great number of non-independent characters in the data, which artificially inflates support for certain groupings and could make ancestral state reconstructions difficult to interpret. We therefore decided not to use binary recoding in our analysis.

As our dataset had many more word meanings than a typical Swadesh list, we were better able to detect semantic shifts. So, we decided to use the etymon-based approach (the first approach) as a first step.[12] All lexical items were put in cognate sets by the Tupí-Guaraní Comparative Project team.[13] Cognate words that had undergone semantic shift were included in their respective cognate

[8] To our knowledge, there is not much more than speculation on the frequency of semantic shift versus loss.

[9] This is not generally true as differences in the semantic boundaries among languages are common. But, many meanings are indeed comparable and stable across languages in general or at least across languages from a certain family due to shared environment, culture, etc.

[10] Typically the more parameters a model has, the better it fits the data. Nevertheless, too many parameters can create an overparameterization problem where the data are not enough to estimate with confidence all the parameters, which results in low accuracy.

[11] This last step makes the resulting matrix very similar to the first type of coding with one crucial difference: the binary recoding ignores presences of an etymon in a language if it has undergone a semantic shift.

[12] We are also planning to code the same dataset in a meaning-based multistate approach and compare the results.

[13] Team members include: Keith Bartholomei, Zachary OHagan, Michael Roberts, and Vivian Wauters, all under the guidance of Professor Lev Michael.

sets.[14] Compound words were temporarily placed only in one cognate set due to time constraints. Each cognate set was then converted to a binary character (1: etymon present in the language, 0: etymon absent). When a language had no entry for a semantic slot, it was coded as unknown for all cognate sets associated with this meaning.

This procedure resulted in a matrix of 3,361 characters with 31.15% missing data. We detected 548 cases of semantic shift out of 10,600 entries or 5.17%.[15] Semantic shift is probably overestimated in the dataset for two reasons. First, there were cases where presumably the same historical semantic shift was counted multiple times (once per daughter language that shows the shift). Second, some of the sources had more generic or vague meanings than others, thus creating "extra" cases of semantic shift. The influence of missing data and semantic shifts on our analysis are further discussed in §4.

2.3 Parsimony Analysis

Parsimony is an intuitive optimality criterion used extensively in both biology and linguistics even before computational methods were developed. It is based on the notion that the simplest explanation (the one that involves the least changes or steps) is the preferable one. One of its main advantages, apart from its simplicity and intuitiveness, is that it does not need an explicit model of evolution and thus can be used on a variety of character types. Parsimony has been shown to be statistically inconsistent under certain conditions (Felsenstein 1978).[16] These conditions, usually described as "long branch attraction," involve long branches (such as language isolates without close relatives in the analysis) and quickly evolving characters that have a limited number of possible states (Schulmeister 2004).

We decided to do a parsimony analysis of our dataset mainly for comparative purposes, as plain parsimony is not the best method to analyze our characters. The reason for that has nothing to do with the "long branch attraction" problem, as we have no language isolates, and although our characters are binary, we do not expect them to evolve quickly (i.e., we do not expect any given etymon to switch many times from present to absent and vice versa). The main problem is that we expect an asymmetry in the rate that cognates are lost versus gained and straight parsimony assigns the same "penalty" to both directions. One could argue that we could weigh the gains more than the losses, but this would involve a more or less arbitrary choice of a number. So, we decided to do a plain parsimony analysis and compare the results with those of the more realistic Bayesian analysis (see below).

As a support measure for our parsimony analysis, we used bootstrapping, a resampling method. For each iteration of the bootstrap algorithm (called a pseudoreplicate), a surrogate dataset equal in size with the original is produced through resampling the characters of the original dataset with replacement. This results in certain characters being omitted and others overrepresented. Then, the same analysis is performed on this surrogate dataset as in the original dataset and the resulting

[14] All the resulting cognate sets are available on request.

[15] The section of kinship terms was not included when calculating the amount of semantic shift because some of the sources were not accurate in their description of the kin system and the amount of semantic shift was overestimated.

[16] A method is statistically consistent if it is guaranteed to approach to the correct solution if given enough (technically infinite) data that are generated according to a specific set of rules (the model).

shorter tree(s) are recorded. This procedure is repeated a large number of times and the results are summarized in a 50% majority tree (where 50% means that a clade/subgroup with 50% bootstrap value was present in 50% of the collection of trees at the end of the bootstrap analysis). A level of 50% is usually chosen as a cut-off point because clades present in more than 50% of the trees are guaranteed to be compatible with each other. There is considerable disagreement and debate on what bootstrap values actually mean in a phylogenetic context and how to interpret them (for more information see §4).

The parsimony analysis was performed in PAUP* 4.0b10 (Swofford 2003). Characters were equally weighted. We performed 40 heuristic searches starting each time with a tree built by stepwise addition of languages and the addition sequence was random. We also performed bootstrapping to calculate support values for our nodes (Felsenstein 1985). We did 1000 pseudoreplicates on our whole dataset and 5000 pseudoreplicates on only the parsimony informative characters, as too many parsimony uninformative characters might artificially decrease support (Soltis and Soltis 2003).

2.4 Bayesian Analysis

The essential outcome of a Bayesian analysis is the posterior probability of the parameters of our model given the data we have. The stereotypical coin-flip equivalent is, how possible is it that the coin is fair (the parameter of our model) given the fact that we just got 90 heads in 100 flips (our data). For a Bayesian analysis, we start with a model of evolution that has some parameters, with our prior beliefs for the probability distribution of these parameters, and with our data, which we assume were generated according to this model.[17] The algorithm then gives us the posterior probability distribution of our parameters and a typical way to summarize the results is to take the 95% confidence interval for each parameter.[18]

For our analysis we chose the restriction site model implemented in MrBayes 3.2.0 (Huelsenbeck et al. 2001; Ronquist and Huelsenbeck 2003). This model allows for two different rates of cognate gain and loss. We believe that this is a more realistic model for our data than a symmetrical model, as we expect that cognate loss is easier than cognate gain. However, it should be noted that no rates were indicated a priori, rather the prior for these two rates is equal and they are able to diverge from there in the expected or the opposite direction. The data were corrected for non-observable all absent sites.[19] We also allowed for variation in the rates across characters, to simulate the situation of quickly evolving etyma (high probability for loss and gain, e.g., through borrowing) and slowly evolving ones. The rate variation across etyma was set to follow a gamma

[17]Of course, if the data are really generated according to our model, then the method is guaranteed to find the correct values of the parameters if given enough data. In reality, though, any evolutionary model is less complex than the real evolutionary process that generated the data. Simulations with generated data under our most sophisticated models and analyzed with our crudest models still give correct results in most cases, showing that phylogenetic methods are robust when the model assumptions are violated (Sullivan and Swofford 2001).

[18]The 95% confidence interval encompasses the most probable values of each parameter given our data (the remaining 5% is the tails of the posterior probability distribution.

[19]This is because we cannot observe etyma that do not exist now in any language because they were gained and then lost from every language. This correction allows the model to be estimated properly by adding dummy characters to compensate for this bias.

shape distribution and the α parameter had a uniform prior from 1 to 200.[20]

Bayesian analysis was conducted with MrBayes 3.2.0. We performed two independent Metropolis-Coupled Markov Chain Monte-Carlo (MCMCMC) runs of 4 chains each (1 cold and 3 heated).[21] The analysis ran for 5 million generations and was sampled every 1,000 generations.

3 Results

3.1 Parsimony

There were 1,162 parsimony informative characters. The heuristic searches resulted in 2 equally parsimonious trees of 4,958 steps. The two trees differed only in the position of Guarayú (sister to Xetá and Tapieté or sister to Paraguayan Guaraní). The strict consensus of the 2 trees is shown in Figure 2. Bootstrap values are visible for nodes with more than 50% support. The bootstrap values for a few branches increased a bit when using only parsimony informative characters, but overall the two bootstrap analyses gave very similar results and no value changed from lower than 50% to higher or vice versa. The numbers shown are from the bootstrap analysis on the parsimony informative characters only.

The parsimony analysis produced a small number of strong groupings supported by a large number of characters (although this support is partly artificial, see §4). All of them except for one are at the tips of the tree and group only 2 or 3 languages together. All the higher level groupings have much less supporting characters (as the lower than 50% bootstrap values show).

[20]The gamma distribution can take a variety of shapes depending on the value of the parameter α.

[21]The samples of the posterior probability distributions are taken only through the cold chain. The 3 heated chains are used to explore tree space more effectively (they can cross areas of low likelihood easier than the cold chain and so they do not get stuck on local optima) and they can swap places with the cold chain if they find an area of higher likelihood.

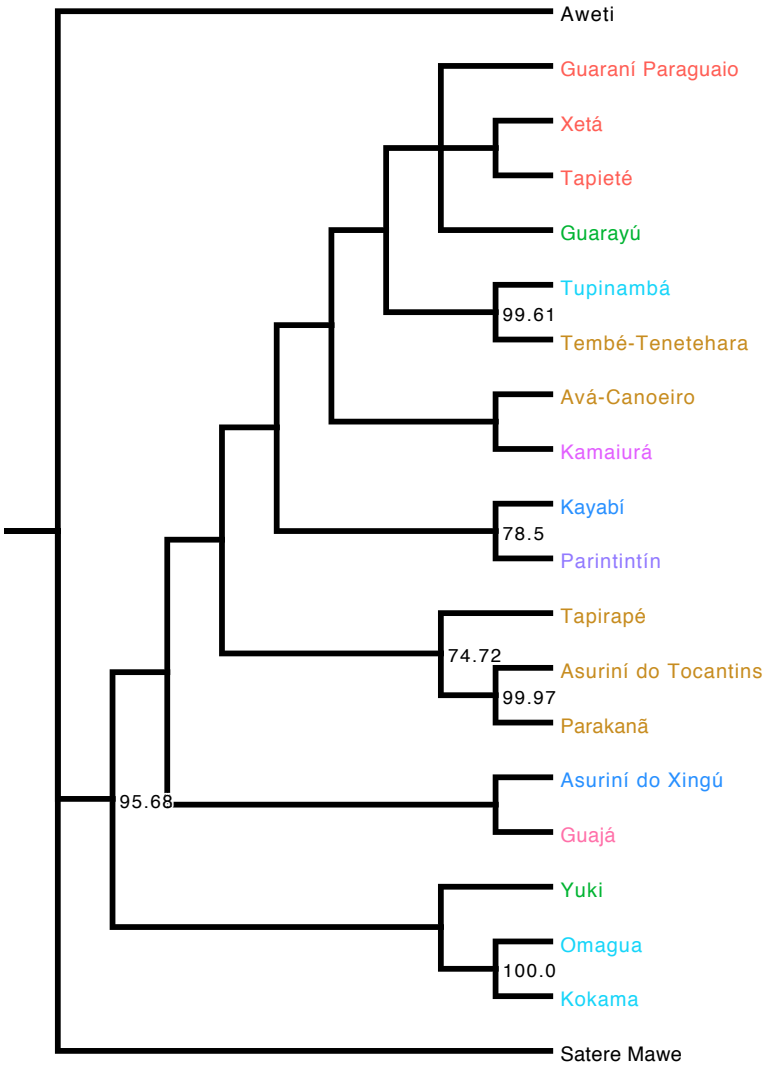

Figure 2: Strict Consensus of two equally parsimonious trees with bootstrap values. The languages are color-coded according to previously proposed groupings. (Values less than 50% not shown.)

3.2 Bayesian Analysis

The majority consensus tree resulting from the Bayesian analysis is shown in Figure 3. The names of the languages are color-coded according to the proposed 8 groups. The values on the branches are the posterior probability of the clade supported by each branch. The branch lengths represent the amount of change that has occurred along the branch and in this value, time and rate are confounded (i.e., a long branch could be due to longer time, fast rate of change, or both). The rates of etymon gain and loss estimated through the analysis were very asymmetrical in the expected direction with the loss rate being 9 times the gain rate.

Convergence between the 2 runs was determined by the average standard deviation of split frequencies (which should approach 0 as convergence is reached, and was in fact 0.0072 at the end of the 5 million generations). The Potential Scale Reduction Factor (PSRF), another convergence

metric, was between 0.999 and 1.001 for all parameters, which shows that the sample for the estimation of each parameter was adequate and the 2 runs had converged. 25% of the initial samples were discarded as burn-in, to ensure that each run had reached stationarity before starting to sample it.

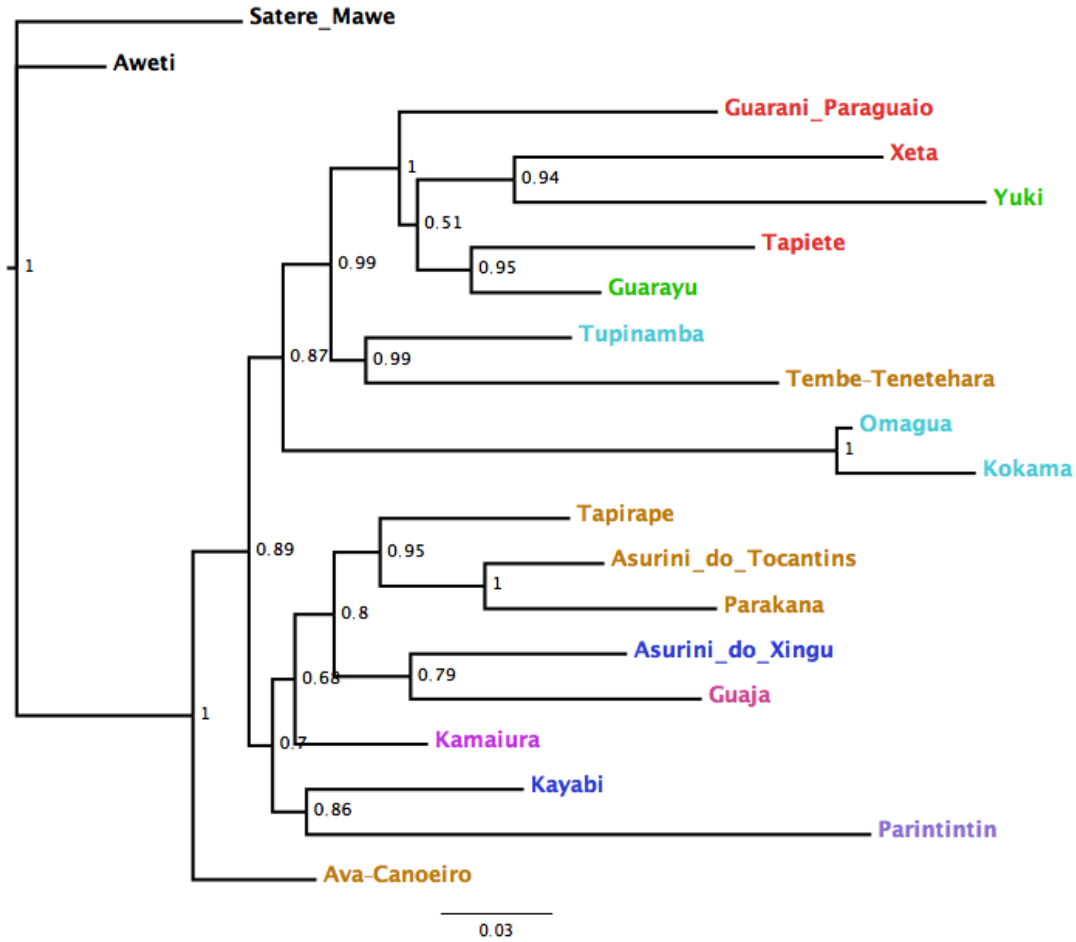

Figure 3: Majority-rule consensus tree with posterior probability values. Languages are color-coded according to previously proposed groupings

4 Discussion

4.1 Characteristics of Dataset

4.1.1 Advantages

Our dataset is the most complete to date for the Tupí-Guaraní family. It contains approximately half of the languages in the Tupí-Guaraní family and it has good coverage of the suggested groups within the family. It also includes 572 word meanings, which is almost 3 times the size of a typical Swadesh list used in other studies (Gray et al. 2009; Atkinson and Gray 2003; Nakhleh et al. 2005). We believe that by including as much information as possible, we have a better chance to recover

subgrouping information for the languages in question. The Swadesh list can be a starting point, but there is no a priori reason to restrict the lexical items collected to the ones included in the list, as the terms that are stable for a language group is to a certain degree idiosyncratic. On the other hand, choosing a few "good" cognate sets to analyze without explicit criteria does not test all the available evidence. As far as the coding method is concerned, we are not aware of another study that used etymon-based coding including reflexes that have undergone semantic shift. Overall, our dataset is both extensive and provides good sampling of the Tupí-Guaraní family and is promising for phylogenetic analyses.

4.1.2 Limitations

There are a number of qualities of our current dataset that limit its usefulness. In order to further develop this analysis in the future, it is important to recognize these limiting factors, and how they contribute to deceptive trends in the data, in order to mitigate deleterious effects in the future. Our two main limitations are: 1) non-independence of the data, and 2) low-quality sources providing misleading data.

Non-independence of our characters means that the states of one character affect the states of other characters. In our dataset, this is caused by languages losing an etymon when they gain another etymon for the same meaning, instead of the erstwhile item shifting to a new meaning. The tendency toward loss instead of shift leads to a high probability that the presence of a reflex in one cognate set generally corresponds to a lack of a reflex in the other cognate sets for that particular item, as seen in Example (1). While we hoped that a high level of shift (i.e., non-loss of words even when a new cognate entered a semantic domain) would mitigate the level of non-independence caused by cognate loss, the level of shift, at 5.17% was not high enough to rule out non-independence as a confounding factor.[22] Additionally, it has been shown for Indo-European that not controlling for non-independence can be deleterious to anything but the smaller subgroupings within the family (Rexová et al. 2003). Nevertheless, the low overall percentage of semantic shift does not capture its distribution over different lexical items. In most lexical items, there was no detected shift at all or just sporadic cases of semantic shift, but in a few cases there was extensive semantic shift that essentially made the etyma independent of one another. Some examples of these semantic "complexes" are: face-head-eye-cheek, nipple-breast-chest, heart-liver-intestines-belly, sand-beach-dust, house-village-shelter, plant-bury-hide, want-love-like. We will explore the possibility of including etymon-based coding only for such items (see §4.4).

(1) 'want'

(1) Paraguayan Guaraní *yvoty*; Tapieté *mbaˈe potɨ*; Guarayú *iˈpotrɨ*; Yuki *bajúti*; Tupinambá *potyra*; Tapirapé *ywãtyr, patyr, hywatyt*; Parakanã *potyr*; *potɨʁa*; Tembé *putir, poti, potyra, iboti, i-putir*; Kayabí *ˈywotyr*; Parintintin *yvatyˈri, yvytyr*; Kamaiurá *potyt*; Guajá *mɨtɨr, mɨtɨ́*; Awetí *potɨt*; Mawé *pohɨt*

(2) Omagua *sɪsa*; Kokama *tsetsa*

[22]This percentage of course refers to our dataset, which included only a fraction of the etyma present in the languages.

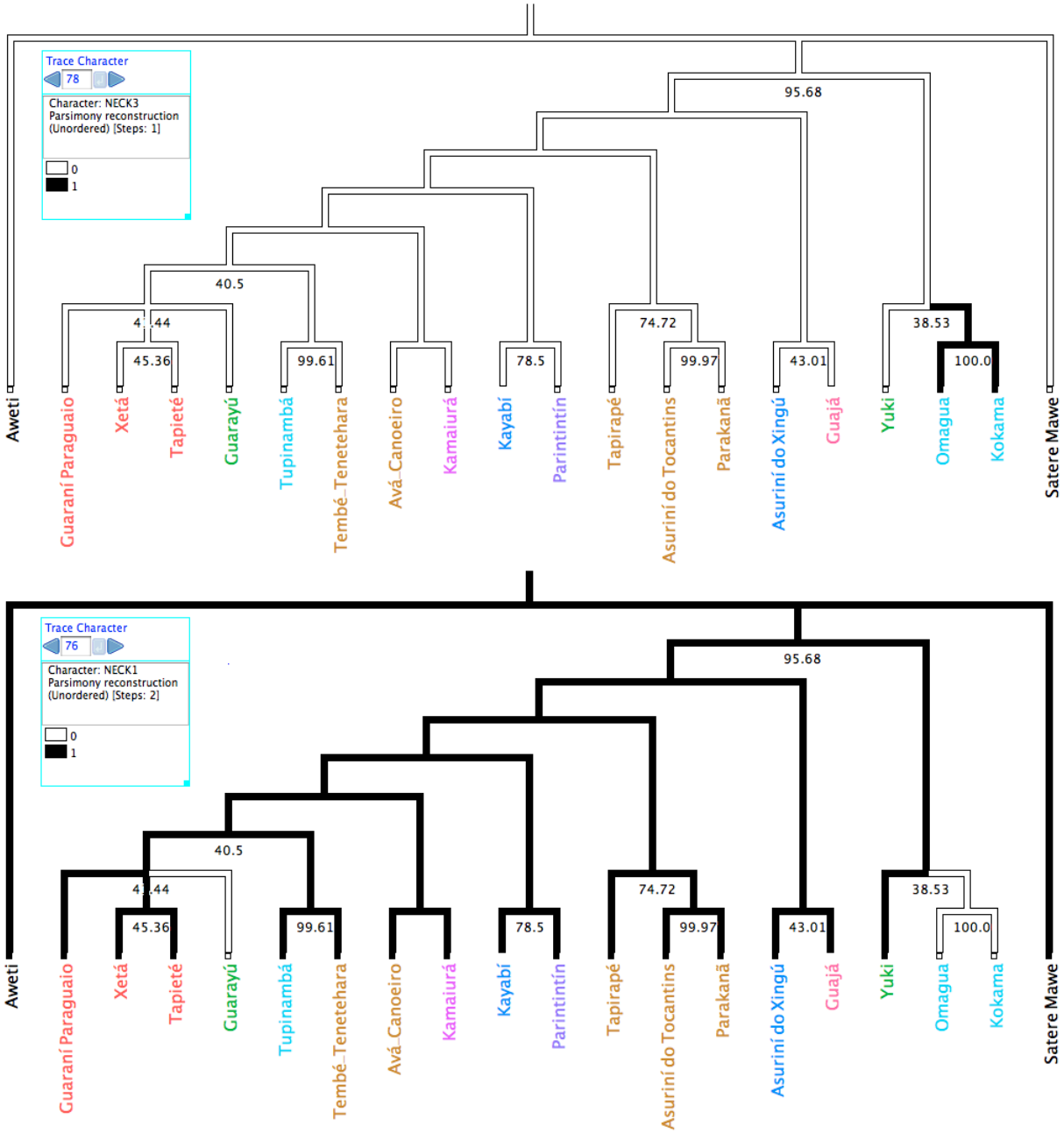

Figure 4: Parsimony ancenstral state reconstruction of cognate sets NECK3 (top) and NECK1 (bottom)

One effect of non-indepedence is that it exaggerates any trends in the data, in the following manner: if two languages exclusively share a cognate set, such as Omagua and Kokama, they are coded the same (as 1) for that set, while all other languages are coded as different (as 0). However, in the following set, where all other languages have a cognate except Omagua and Kokama, they will be coded exactly opposite, thus creating *two* data points that link Omagua and Kokama as an exclusive pair to the exclusion of the rest of the languages. Some cases of trends like this are shown in Figure 4 as well as Example (1) using the words for "neck" and "want," respectively. In order

to argue for the independence of these two points, we would be positing that the loss of a reflex of one cognate and the presence of a different cognate for the same lexical item are completely unrelated. Since this is not the case, it may have artificially increased the support for subgroupings that have a high number of shared innovations. In our current dataset, this is most likely the case for Omagua and Kokama because of the high level of non-Tupí-Guaraní lexical items in their language. However, this could also be contributing to the high numbers for the Parakanã, Tapirapé, and Asuriní do Tocantins subgrouping.

Another effect of non-independence is realized in the marking of unknown character states. As described above, for each lexical item that we did not find an entry for in a particular language, we coded each character (etymon) associated with this semantic slot as unknown, with a question mark ('?'). In this way, the analyses treated these unknown cases as either present or absent, in order to fit the data. The number of unknown states for a given character was entirely dependent on how many known cognates for that character there were in the other languages. For example, our data on Xetá lacks both a word for "flute" and "bead." However, there are 12 different cognate sets for "flute," and only one for "bead," which result in a total of 13 unknown character states in Xetá. It is still unclear what the final effect of these characters is on our analysis.

Another possible distortion from the data stems from variability in data quality. Each of the sources has a different level of detail in differentiating the semantic domain of each lexical item. In particular, the length of the branches in MrBayes could be distorted by an overrepresentation of single set cognates in a single language for a single lexical item. This is most evident in Parintintin, which is represented with a very long branch length; the source for Paritintin is a dictionary that makes very little subcategorization of terms in the dictionary. For example, the lexical entry "bee" has 14 entries, none of which are cognate, the lexical for "cousin" (gender-unspecified) has 8 non-cognate entries, and the lexical entry "road" has 6 non-cognate entries. This creates an abundance of single-item cognate sets for Parintintin, which leads MrBayes to analyze it as extremely divergent compared to other languages. This is also the case for Tembé, which has for instance 8 non-cognate words under the lexical item "beat." While one would expect that each of the words listed under a single heading actually have some semantic distinctions between them, the lack of description in the literature causes a spurious effect in the data.[23]

4.2 Comparison of Methods

The parsimony analysis recovered a few low-level subgroups supported by a large number of characters, as well as the Tupí-Guaraní family. All other groupings had low bootstrap values. As mentioned in the methods section (§2), bootstrapping has been used as a measure of support or even accuracy (Hillis and Bull 1993), but there is considerable debate over its interpretation (Soltis and Soltis 2003). The only thing the low bootstrap values tell us for sure about a branch is that there are few characters supporting this branch within our dataset. This does not necessarily mean that the clade is not accurate. In fact, if the branch is short (because the period of shared history for the clade was short), it is very probable that we will have few characters that change along this

[23]Note, however, that the branch length of Yuki does not appear to be an artifact of under-informative data, but rather the Yuki lexicon is remarkably divergent, having lost many of the Tupí-Guaraní cognates found in many other languages in the family.

branch. In this case, we will never get a high bootstrap value, even if there is nothing else that contradicts this grouping. So, bootstrap values are good indicators of how much support there is in the dataset for a particular grouping and if there are characters that support conflicting groupings.

Both the parsimony and the Bayesian analyses recovered the same highly supported low-level subgroups (Tupinambá and Tembé, Omagua and Kokama, Tapirapé sister to Parakanã and Asuriní do Tocantins), as well as the Tupí-Guaraní family node. However, the Bayesian analysis showed two additional well-supported subgroups: the clade including all languages from Groups 1 and 2 (although its internal relationships are not fully resolved) and its sister relationship with the Tupinambá-Tembé clade. All of these languages were grouped together in the parsimony analysis, but with low support, except for Yuki which was grouped with the Omagua-Kokama clade.[24]

As mentioned in §2, we consider the model used in the Bayesian analysis much more realistic than the symmetrical penalty of the parsimony analysis. The rates of cognate gain and loss estimated through the Bayesian analysis confirmed our expectations, as they were highly asymmetrical. Also, the Bayesian analysis is able to use all the coded characters (while parsimony was using only a third of them, which were parsimony informative). Therefore, we consider the results of the Bayesian analysis to be more trustworthy with this type of data. However, even with the better model, the posterior probability of higher-level subgroupings in the Bayesian analysis is still low, in many cases much lower than the 95% confidence interval. Many higher-level branches are also fairly short. It could be that the Tupí-Guaraní languages diversified fast and there is no adequate evidence to resolve the topology of the tree at that level. Additional data from other lines of evidence (phonology, morphology) could help clarify these relationships (see §4.4). In case the base of the Tupí-Guaraní family is a real polytomy, with many subgroups diversifying more or less at the same time, phylogenetic reconstructions will be sensitive to the addition and removal of characters, as there will be many conflicting characters for different topologies.

4.3 Promising Trends

Despite the possible confounds to our analysis, this preliminary study is demonstrably useful based both on a number of expected subgroupings based on previous reconstructions, as well as particular structures that point to revelatory possibilities for re-analysis of the relationships between specific languages.

4.3.1 Comparison with Previous Reconstructions

Because a full understanding of the accuracy and practicability of using phylogenetic methods on linguistic data has not been conclusively determined, one of the goals of recent phylogenetic studies has been to evaluate the results gathered in comparison to previous research (Dunn et al. 2008; Atkinson and Gray 2003). In order to determine the basic validity of our data and methods, we used two outgroup languages, as discussed above (§2). We rooted the tree using Mawé, since it is considered to be further outside the family than Awetí, and then our analysis of the data as a family was validated by the unequivocal placement of Awetí as outside of the Tupí-Guaraní family.

[24]Yuki is a highly divergent language and we may need additional and more reliable characters to place it with confidence.

This shows that the family is monophyletic, i.e., that the languages can be reliably determined to be a family to the exclusion of the two outgroup languages.

Having determined that our data did in fact produce an exclusive Tupí-Guaraní clade, it is possible to compare the results of our current analysis with the previous subgrouping hypotheses. As discussed above, the diachronic study of the Tupí-Guaraní family has, for the most part, not been rigorously undertaken, and therefore the 8 conventional groups into which languages have been divided is questionable (of course, Mello 2000 does not follow this convention entirely). Nevertheless we will discuss the points of agreement of our analysis with the 8 groups, and then separately discuss important comparisons with Mello (2000) not covered by the prior hypotheses.

In the parsimony analysis as well as MrBayes, the most consistent groupings reproduced above the pairwise level (i.e., more than two languages put together) were in Groups 1 and 4. For the parsimony analysis, the three Group 1 languages were reproduced in a clade with Guarayú, but the internal topology of the clade was unresolved. This clade is a combined Group 1 and Group 2 clade with the exception of Yuki.

In MrBayes, Guarayú and Yuki were both included with the Group 1 languages. In the case of Group 4, both trees produced an identical subgrouping of three languages: Parakanã with Asuriní do Tocantins, followed by an immediately higher node of Tapirapé. Both trees did not include Tembé in this clade, and instead clustered Tembé with Tupinambá, the possible reason for which is discussed above. Also reproduced in these trees was the close relationship between Omagua and Kokama, which, in spite of the problems discussed above, is an encouraging result.

In contrast to the congruence between the current analysis and those of Jensen (1998), the Mello (2000) language groupings are not as well-matched with our analysis, partially due to differences in which languages were included in each sample. Paraguayan Guaraní and Xetá, which form a subgroup in Group 1, do appear as part of a larger clade including Guarayú, which is in Group 3 for Mello (2000). Also, while the Group 6 languages (Group 4 in Jensen 1998) Parakanã, Asuriní do Tocantins, and Tapirapé pattern together perfectly, even following the subgrouping in Mello (2000) with Tapirapé appearing one node higher, Asuriní do Xingu, which is included in this Group by Mello (2000), is not reproduced in either the parsimony analysis or MrBayes.

While Dietrich (1990) explicitly states that his work is not a subgrouping analysis, the gradient conversativeness rating of each language does result in some low-level groupings, many of which follow Jensen (1998). There are, however, some important differences. Dietrich does not find Tupinambá and Kokama to be similar enough for subgrouping. Furthermore, like Mello (2000), the conservativeness ranking finds Asuriní do Tocantins to be an outlier in a Group containing Guajajára and Tembé (which we consider dialectal variants of the same language) as well as Kamaiurá, a subgrouping that is not supported by either of our analyses. Further, Dietrich finds Kayabí to be much more closely related to Tapirapé and Kamaiurá than to Asuriní do Xingu. This is another subgrouping that is not supported by our analyses, although both the parsimony and Bayesian analyses split Kayabí and Ausuriní do Xingu into very different trees. This points to considerable confusion about the placement of Kayabí, as it is one of the significant differences between Jensen (1998) and Rodrigues and Cabral (2002) (see §1.2).

In summary, there is a reasonable level of agreement between our analysis and that of Jensen (1998) about some basic groupings in the Tupí-Guaraní family, and slightly less agreement with Mello (2000). However, there were a number of distinctly different subgroupings. Both the parsi-

mony analysis and MrBayes placed Yuki in different positions (with Omagua and Kokama for the parsimony analysis and with Xetá for MrBayes), neither of which matches any previous analysis. As discussed above, we attribute this to the highly divergent lexicon of Yuki. Because non-cognate forms were all coded the same (as 0), the analyses tended to group divergent languages together. However, both parsimony and MrBayes grouped Guajá with Asuriní do Xingu, which is not supported by any previous analysis. This particular grouping will be more informative once data on the other languages in our sample have been gathered because currently only one language from each of the previously hypothesized subgroups to which these languages belong has been included in the sample. Therefore, it may be that the two are grouping together based on a lack of more closely related languages, and the inclusion of other languages in Group 6 and 8 will result in a different pattern in our results. If this is not the case, however, it may indicate that the languages in these two groups, like those in Groups 1 and 2, may be better analyzed as all part of a larger clade within the Tupí-Guaraní family.

4.3.2 Subgrouping Implications

The original purpose of the Tupí-Guaraní Comparative Project was to correctly determine the relationship of Omagua to the Tupí-Guaraní family. Based on this research goal, the most promising result from this analysis is the way in which the languages of the erstwhile Groups 1 and 3 combine. The inclusion of Tupinambá and the Omagua-Kokama clade as successive branches of the Group 1 (more or less) clade points to a need for re-analysis regarding the immediate proto-language of Omagua and Kokama, not as the direct descendent of Tupinambá, as previously suggested (Cabral 1995). In previous research, evidence has focused on the relationship of Kokama (and thus Omagua) as Tupí-Guaraní or not, always in relation to Tupinambá as the potential predecessor (Cabral 1995). Based on the results of the phylogenetic analysis, however, Omagua and Kokama are very clearly a subgroup, as are Tupinambá and Tembé (although recall that some data issues may be creating a spurious subgrouping in that case), but Tupinambá and the Omagua-Kokama clade are members of a paraphyletic group, in that they do not subgroup to the exclusion of other languages. Instead, the MrBayes analysis suggests that Paraguayan Guaraní and and the other languages in Groups 1 and 2, as well as Tupinambá and the Omagua-Kokama subgroup all share a common ancestor. If this subgrouping turns out to be robust, it would suggest that the Tupí-Guaraní ancestor of Omagua and Kokama is fairly old, which would help explain the large amount of change the language has undergone, changes so dramatic as to make some believe that it is not a Tupí-Guaraní language at all, as discussed above.

4.4 Future Directions

In order to improve the phylogenetic analyses of subgrouping in the Tupí-Guaraní languages, and to further determine the validity of observed trends within our extant dataset, we will include more languages in our dataset and we want to use a variety of characters for phylogenetic reconstruction.

In terms of language addition, we will be adding 7 more Tupí-Guaraní languages as well as at least one more outgroup language from the Tupian stock. This way we will have at least 2 languages per proposed group so that we can test previous subgrouping hypotheses. We will also

be able to test the sister relationship of Awetí with Tupí-Guaraní.

In terms of characters and coding, we are planning to use the meaning-based method of character coding for our lexical characters to avoid non-independence problems. An additional possibility we are going to explore is keeping the etymon-based coding only for the semantic "complexes" that exhibited a lot of semantic shifts in order not to lose any subgrouping information. Apart from coding lexical characters per se, we will use our cognate sets to find and code phonological characters as well. As phonological characters are the most reliable line of evidence in the Comparative Method, it is important to include this type of data in our phylogenetic analyses.

Additionally, we plan to undertake the same analyses discussed in this study on a sample of morphosyntactic functional items in the Tupí-Guaraní languages. This database will include data from the same languages included in the lexical database, but instead of gathering forms based on an established set of items, we have chosen to cull all available functional morphemes from each language, and then organize them by type. The result of this more holistic data collection technique is that we can be more confident in obtaining a full sample of the forms within any given language. As is the case with phonological characters (discussed above), functional items are less prone to superficial effects like contact (Thomason 2001). Therefore, one would expect that any true trends in our data would be more likely to manifest themselves in an analysis of less volatile data, such as the functional items. Furthermore, because functional morphemes can follow grammaticalization trajectories, and might therefore be less likely to disappear from a language due to the entrance of a new morpheme, our assumptions of independence violated above may not be as problematic for the functional list.

In addition to our extension of the current cognacy analysis to functional items and the other types of coding based on semantics and phonology, we will also be reconstructing the Tupí-Guaraní family using the traditional Comparative Method. Despite the valuable insights garnered from phylogenetics, there is currently no viable alternative to the Comparative Methods for achieving respected and believable results. Therefore, our "by hand" reconstruction, as discussed in the introduction, will be invaluable. It will be an indispensable tool for ancestral state reconstruction, as well as providing us with a valuable basis on which to calculate the overall utility and accuracy of these exciting new methods.

References

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

Armoye, Celso. 2009. Análisis de la lengua guarayo (tesina).

Atkinson, Quentin D. and R. D. Gray. 2003. Language-Tree Divergence Times Support the Anatolian Theory of Indo-European Origin. *Nature* 435–438.

Betts, La Vera. 1981. *Dicionário parintintín-português português-parintintín*. Cuiabá: Summer Institute of Linguistics.

Borella, Cristina de Cássia. 2000. Aspectos morfossintáticos da língua awetí (tupí). Ph.D. thesis, Universidade Estadual de Campinas.

Borges, Mônica Veloso. 2006. Aspectos fonológicos e morfossintáticos da língua avá-canoeiro (tupi-guarani). Doctoral thesis, Universidade Estadual de Campinas.

Borges, Mônica Veloso. 2007. *Posposições da língua avá-canoeiro (tupi-guarani)*, 385–389. Campinas: Editora Curt Nimuendajú.

Boudin, Max H. 1978. *Dicionário de tupí moderno: Dialeto tembé-ténêtehar do alto rio Gurupi. 2 vols.* São Paulo: Conselho Estadual de Artes e Ciências Humanas.

Cabral, Ana Suelly Arruda Camara. 1995. Contact-Induced Language Change in the Western Amazon: The Non-Genetic Origin of the Kokama Language. Ph.D. thesis, University of Pittsburgh.

Cunha, Péricles. 1987. Análise fonêmica preliminar da língua guajá. Master's thesis, Universidade Estadual de Campinas.

Derbyshire, Desmond C. 1994. Clause Subordination and Nominalization in Tupi-Guaranian and Cariban Languages. *Revista Latinoamericana de Estudios Etnolingüísticos* 8:179–198.

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

Dixon, R. M. W. and Alexandra Y. Aikhenvald. 1999. *The Amazonian Languages*. Cambridge University Press.

Dobson, Rose M. 1988. Aspectos da língua kayabí. *Serie lingüística* 12.

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

Drude, Sebastian. 2011. *Awetí in Relation with Kamayurá: the Two Tupian Languages of the Upper Xingu*, 155–192. Rio de Janeiro: Museu do índio/FUNAL.

Dunn, M. 2009. Contact and Phylogeny in Island Melanesia. *Lingua* 1664–1678.

Dunn, Michael, Stephen Levinson, Eva Lindström, and Ger Reesink. 2008. Structural Phylogeny in Historical Linguistics: Methodological Explorations Applied in Island Melansia. *Language* 84(4):710–759.

Espinosa, Lucas. 1935. *Los tupí del Oriente peruano*. Madrid: Imprenta de Librería y Casa Editorial Hernando (S.A.).

Etnolingüística. 2011. Tupí-Guaraní. URL `http://www.etnolinguistica.org/familia:tupi-guarani`, [Online; accessed 2-May-2011].

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

Faust, Norma. 1971. *Cocama Clause Types*, 73–105. Summer Institute of Linguistics Publications in Linguistics and Related Fields, Summer Institute of Linguistics.

Felsenstein, Joseph. 1978. Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading. *Systematic Biology* 27(4):401–410.

Felsenstein, Joseph. 1985. Confidence Limits on Phylogenies: an Approach Using the Bootstrap. *Evolution* 39(4):783–791.

Franceschini, Dulce. 2000. La langue sateré-mawé: Description et analyse morphosyntaxique. Ph.D. thesis, ANRT, Lille.

González, Hebe Alicia. 2005. A Grammar of Tapiete (Tupi-Guarani). Doctoral thesis, University of Pittsburgh.

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

Gray, R.D., A.J. Drummond, and S.J. Greenhill. 2009. Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement. *Science* 323(5913):479.

Greenhill, S.J. and R.D. Gray. 2005. Testing Population Dispersal Hypotheses: Pacific Settlement, Phylogenetic Trees and Austronesian Languages. *Mace et al* 31–52.

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

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

Harrison, Carl H. 2009. Pedagogical Information and Drills for the Asuriní Language .

Heggarty, P. 2006. Interdisciplinary Indiscipline? Can Phylogenetic Methods Meaningfully Be Applied to Language Data and to Dating Language. *Phylogenetic Methods and the Prehistory of Languages* 183.

Hillis, David M. and James J. Bull. 1993. An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis. *Systematic Biology* 42(2):182–192.

Huelsenbeck, John P., Fredrik Ronquist, Rasmus Nielsen, and Jonathan P. Bollback. 2001. Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology. *Science* 294(5550):2310–2314.

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

Jensen, Cheryl Joyce. 1998. *Comparative Tupí-Guaraní Morphosyntax*, volume 4, 489–618. New York: Mouton de Gruyter.

Lathrap, Donald W. 1970. *The Upper Amazon*. London: Thames and Hudson.

Lemle, Miriam. 1971. *Internal Classification of the Tupí-Guaraní Linguistic Family*, 107–129. Summer Institute of Linguistics Publications in Linguistics and Related Fields, Summer Institute of Linguistics.

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

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). Doctoral thesis, Universidade de Brasília.

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

Michael, Lev D. 2010. The Pre-Columbian Origin of a Diachronic Orphan: The Case of Omagua.

Nakhleh, L., T. Warnow, D. Ringe, and S.N. Evans. 2005. A Comparison of Phylogenetic Reconstruction Methods on an Indo-European Dataset. *Transactions of the Philological Society* 103(2):171–192.

Nascimento, Ana Paula Lion Mamede. 2008. Estudo fonético e fonológico da língua guajá. Master's thesis, Universidade de Brasília.

Nichols, J. and T. Warnow. 2008. Tutorial on Computational Linguistic Phylogeny. *Language and Linguistics Compass* 2(5):760–820.

Nicholson, Velda C. 1982. Breve estudo da língua asuriní do Xingú. *Ensaios lingüísticos* 5.

O'Hagan, Zachary, Lev D. Michael, Clare Sandy, Tammy Stark, and Vivian Wauters. 2011. Omagua-Spanish-English Dictionary. Ms. Project-internal publication of the Omagua Documentation Project.

Olson, Gary Paul. 1978. Descrição preliminar de orações waiãpi. *Ensaios lingüísticos* 3.

Pease, Helen. 1968. Parintintin Grammar.

Pereira, Antônia Alves. 2009. Estudo morfossintático do asuriní do Xingú. Doctoral thesis, Universidade Estadual de Campinas.

Praça, Walkíria Neiva. 2007. Morfossintaxe da língua tapirapé. Doctoral thesis, Universidade de Brasília.

Rexová, Kateřina, Yvonne Bastin, and Daniel Frynta. 2006. Cladistic Analysis of Bantu Languages: a New Tree Based on Combined Lexical and Grammatical Data. *Naturwissenschaften* 93(4):189–194.

Rexová, Kateřina, Daniel Frynta, and Jan Zrzavý. 2003. Cladistic Analysis of Languages: Indo-European Classification Based on Lexicostatistical Data. *Cladistics* 19(2):120 – 127.

Rodrigues, Aryon Dall'Igna. 1958. Classification of Tupí-Guaraní. *International Journal of American Linguistics* 24(3):231–234.

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

Rodrigues, Aryon Dall'Igna. 2007. *As consoantes do proto-tupí*, 167–203. Campinas: Editora Curt Nimuendajú.

Rodrigues, Aryon Dall'Igna and Ana Suelly Arruda Camara Cabral. 2002. *Revendo a classificação interna da família tupí-guaraní*, 327–337. Belém: Editora Universitária, Universidade Federal do Pará.

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

Sampaio, Wany Bernardette de Araújo. 1977. Estudo comparativo sincrônico entre o parintintin (tenharim) e o uru-eu-uau-uau (amondawa): Contribuições para uma revisão na classificação das línguas tupí-kawahib. Master's thesis, Universidade Estadual de Campinas.

Schleicher, Charles Owen. 1998. Comparative and Internal Reconstruction of Proto-Tupí-Guaraní. Doctoral thesis, University of Wisconsin, Madison.

Schulmeister, Susanne. 2004. Inconsistency of Maximum Parsimony Revisited. *Systematic Biology* 53(4):521–528.

Seki, Lucy. 1982. Marcadores de pessoa do verbo kamaiurá. *Cadernos de estudos lingüísticos* 3:22–40.

Seki, Lucy. 1983. Observações sobre variacão sociolingüística em kamaiurá. *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*, 367–391. Austin: University of Texas Press.

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

Seki, Lucy. 2007. *Partículas e tipos de discurso em kamaiurá*, 145–157. Caracas: Universidad Católica Andrés Bello.

Corrêa da Silva, Beatriz Carretta. 2010. Mawé/awetí/tupí-guaraní: Relações lingüísticas e impicações históricas. Ph.D. thesis, Universidade de Brasília.

da Silva, Gino Ferreira. 2003. Construindo um dicionário parakanã-português. Master's thesis, Universidade Federal do Pará.

Soares, Marilia Facó and Yonne Leite. 1991. *Vowel Shift in the Tupí-Guaraní Language Family: A Typological Approach*, 36–53. Philadelphia: University of Pennsylvania Press.

Soltis, Pamela S. and Douglas E. Soltis. 2003. Applying the Bootstrap in Phylogeny Reconstruction. *Statistical Science* 18(2):256–267.

Sullivan, Jack and David L. Swofford. 2001. Should We Use Model-Based Methods for Phylogenetic Inference When We Know That Assumptions About Among-Site Rate Variation and Nucleotide Substitution Pattern Are Violated? *Systematic Biology* 50(5):723–729.

Swofford, D.L. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4 .

Thomason, Sarah. 2001. *Contact-Induced Typological Change*, 1640–1648. Berlin & New York: Walter de Grutyer.

Vallejos, Rosa. 2010. A Grammar of Kokama-Kokamilla. Doctoral thesis, University of Oregon.

Vasconcelos, Eduardo Alves. 2008. Aspectos fonológicos da língua xetá. Master's thesis, Universidade de Brasília.

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

Natalia Chousou-Polydouri
University of California, Berkeley
Department of Environmental Science, Policy, and Management
130 Mulford Hall # 3114
Berkeley, CA 94720
nataliacp@berkeley.edu

Vivian Wauters
University of California, Berkeley
Department of Linguistics
1203 Dwinelle Hall
Berkeley, CA 94720
vmw@berkeley.edu

REPORT 15

SURVEY OF CALIFORNIA AND OTHER INDIAN LANGUAGES

Structure and Contact in Languages of the Americas

John Sylak-Glassman and Justin Spence, Editors

Andrew Garrett and Leanne Hinton, Series Editors

cover design by Leanne Hinton (Santa Barbara Chumash rock painting)

Table of Contents