A Test of Coding Procedures for Lexical Data with Tupí-Guaraní and Chapacuran Languages

Natalia Chousou-Polydouri,* Joshua Birchall,† Sérgio Meira,† Zachary O'Hagan,‡ and Lev Michael‡
*Laboratoire Dynamique du Langage
†Museu Paraense Emílio Goeldi
‡University of California, Berkeley

***Abstract*—Recent phylogenetic studies in historical linguistics have focused on lexical data. However, the way that such data are coded into characters for phylogenetic analysis has been approached in different ways, without investigating how coding methods may affect the results. In this paper, we compare three different coding methods for lexical data (multistate meaning-based characters, binary root-meaning characters, and binary cognate characters) in a Bayesian framework, using data from the Tupí-Guaraní and Chapacuran language families as case studies. We show that, contrary to prior expectations, different coding methods can have a significant impact on the topology of the resulting trees.**

***Keywords*—Bayesian phylogenetic inference, cognate coding, historical linguistics, South American indigenous languages**

I. Introduction

South America, long considered the ethnographically and linguistically "least known continent" [1], has in recent decades experienced a surge of descriptive and documentary linguistic research [2], [3]. The classification of the languages of this region, and especially those of Amazonia, has, in contrast, advanced little in the last 50 years [4], [5]. However, the increasing availability of lexical data on South American languages, as well as recent successes in applying computational phylogenetic techniques to data of this type, offers us the opportunity to push forward our understanding of genealogical relationships in the region with new datasets and tools [6]–[8].

While it is accepted that lexical data from natural languages carry phylogenetic signal, the study of lexical evolution per se has largely been neglected by historical linguistics (with the exception of lexicostatistics), as the evolution of other domains of language, such as phonology and morphology, are considered more informative for subgrouping and less susceptible to borrowing. In contrast, computational phylogenetic studies in recent years have focused primarily on lexical evolution, due to the ease with which relatively short wordlists can be analyzed with a variety of established phylogenetic methods.

A critical aspect of these methods, and a way in which they differ, is the manner in which phylogenetic characters are generated from lexical data. The differing nature of these characters ultimately reflects different understandings of the phylogenetic notion of homology [9] in the context of lexical evolution. However, there has been little discussion of the implications of different coding methods and what the underlying assumptions of each are regarding how the lexicon evolves. At the same time, there has been little work to evaluate if and how different coding methods affect resulting classifications, with two exceptions: a parsimony-based empirical test on Indo-European by Rexová and colleagues [10] and an analytical investigation of Pagel and Meade based on a maximum likelihood framework [11]. While Rexová and colleagues find topological differences when using different coding methods, Pagel and Meade predict no impact on topology, although differences in branch lengths and support values are expected.

In this paper, we briefly describe and discuss three major lexical coding methods and we compare their results in a Bayesian Inference framework, using data from the Tupí-Guaraní and Chapacuran language families as case studies.

II. Data

We test the different coding methods on lexical datasets for two South American language families: a Tupí-Guaraní dataset of 33 languages for a 547-meaning wordlist [7], and a Chapacuran dataset of 11 languages for a 126-meaning wordlist [8]. Each dataset includes data for every language for which adequate lexical data is available.

III. Methods

A. Coding procedures

We compare three coding procedures based on different types of characters: 1) multistate meaning-based characters; 2) binary root-meaning characters; and 3) binary cognate characters. The two first coding methods are based on a comparative lexical dataset collected using a wordlist, while the third necessitates the broader collection of lexical data including close synonyms.

A typical comparative lexical dataset based on a wordlist yields inherently multistate characters. Each meaning of the wordlist is a character. All languages that exhibit cognate forms for a given meaning are given the same character state value. In other words, each character is equivalent to the question "For meaning X, what root (or roots) express X?" and the coding method essentially tracks lexical replacement. We refer to this scheme as 'multistate meaning-based' coding. Surprisingly, this coding method has been very rarely used [10]. Among its advantages is the ease of data collection and its applicability in instances of little available lexical data. One potential problem of multistate meaning-based coding is that it

can be affected by parallel semantic shift, which would make the same state arise multiple times independently (see [12]).

Most phylogenetic studies perform an additional binary recoding operation on the above-described multistate characters before using them as input into the analysis [13], [14]. This operation converts each character state of the multistate characters into a binary presence-absence character, which we refer to as a 'binary root-meaning' character, following Chang et al. [12]. This conversion essentially yields a dataset that answers the question "Does a given language have a reflex of a particular ancestral form A that means X?" Binary root-meaning coding shares all the advantages of applicability and ease of data collection with multistate meaning-based coding, in addition yielding binary characters which are easy to model and treat computationally.

However, despite the superficial resemblance of root-meaning characters to cognate sets, root-meaning coding poses serious conceptual problems. First of all, root-meaning sets often constitute only partial cognate sets, since they do not contain cognate words that have undergone semantic shift. More importantly, however, they subsume under the same state (absence) the results of two very different evolutionary events: root loss and semantic shift, both of which can 'break' a root-meaning association. This can potentially lead to spurious subgroups supported only by such 'shared' absences (which in fact do not correspond to a shared evolutionary event). Furthermore, binary root-meaning coding creates characters that are not independent, violating a key assumption of all phylogenetic methods [15]. This can be seen by considering the following example. If a language family has three roots associated with the meaning 'head' (i.e., HEAD-A, HEAD-B, and HEAD-C), and a particular language exhibits a form belonging to HEAD-A for the meaning 'head', we can predict that it lacks forms belonging to HEAD-B and HEAD-C that also mean 'head' (except in the rare case where there are two exactly synonymous forms and both of them have been collected as the most basic form). Of course, in reality the language may very well have forms that are cognate with HEAD-B and HEAD-C, but they may have undergone semantic shift. Finally, binary root-meaning coding is also susceptible to parallel semantic shift. The multistate meaning-based and binary root-meaning coding schemes are exemplified in Table I with the Chapacuran forms for 'wing' and 'feather'.

The third coding scheme uses cognate sets as the basis for binary presence-absence characters, defined as a set of forms with a common ancestral form, or etymon [7], [8]. Characters of this type conform to the traditional notion of cognacy in historical linguistics, since they include forms that have undergone semantic shift. Although this coding method is more difficult to apply, as it requires relatively extensive lexical sources and etymological knowledge, it largely avoids the non-independence problem and parallel semantic shifts are not treated as shared evolutionary events. Cognate set coding is exemplified in Table II, where the same Chapacuran forms for 'wing' and 'feather' are coded as cognates that have undergone semantic shift.

TABLE I
MULTISTATE MEANING-BASED AND BINARY ROOT-MEANING CODING

	ORO WIN	WANYAM	JARÚ	MORÉ	KITEMOKA
Wing	napat	nipat	tinji	nipat	?
Feather	tyne	nipat	?	tain	ipati
WING	A	A	B	A	?
FEATHER	A	B	?	A	B
WING-A	1	1	0	1	?
WING-B	0	0	1	0	?
FEATHER-A	1	0	?	1	0
FEATHER-B	0	1	?	0	1

TABLE II
BINARY COGNATE CODING

COGNATE	ORO WIN	WANYAM	JARÚ	MORÉ	KITEMOKA
Wing	napat	nipat	tinji	nipat	?
Feather	tyne	nipat	?	tain	ipati
WING1	1	1	0	1	1
FEATHER1	1	0	1	1	0

B. Bayesian phylogenetic inference

We use a simple binary model for the binary characters and the Mk model for the multistate characters [16]. The analysis of the Tupí-Guaraní data was done in MrBayes 3.2 [17], [18] and the trees were rooted with Mawé as the outgroup, following the current consensus in Tupian studies [19], [20]. The analysis of the Chapacuran data was done in BEAST v.1.8.2. [21] using a constant size coalescent tree prior, a strict clock and tip-dates. For more detailed information regarding phylogenetic analyses, see [7], [8].

IV. RESULTS

For the results and discussion, we use a significance cutoff value of 0.80 posterior probability. In all figures, the highlighted branches support clades that are unique to a particular coding method. The rest of the branches support clades that are either poorly supported or common among at least two coding methods.

A. Tupí-Guaraní dataset

For the Tupí-Guaraní dataset, different coding methods produce drastically different results, both in regards to topology and posterior probabilities. The multistate meaning-based coding (results not shown) produced results largely compatible with the binary root-meaning coding, but the latter had much higher posterior probabilities and therefore resolution. On the other hand, the topological differences between the binary root-meaning coding and the cognate coding are striking, with a large number of unique nodes present in each tree (see Figures 1 and 2). Overall, the cognate coding method produced a more resolved tree than the root-meaning coding (25 vs. 17 nodes supported above the cutoff), and recovered a greater number of the eight traditionally recognized subgroups [22] (five vs. four). Many of the unique nodes supported by

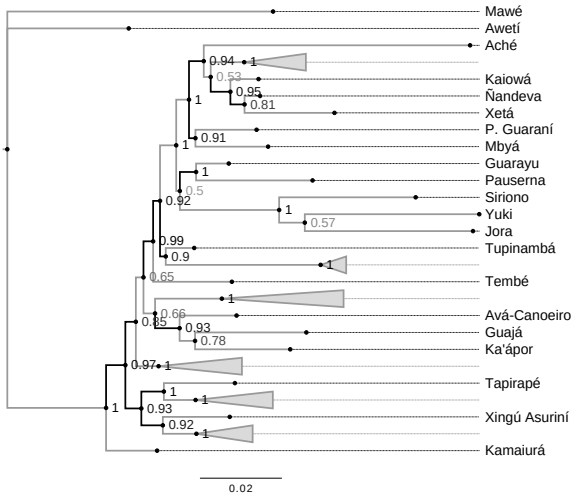

Fig. 1. Tupí-Guaraní: Majority-rule consensus tree of binary cognate coding

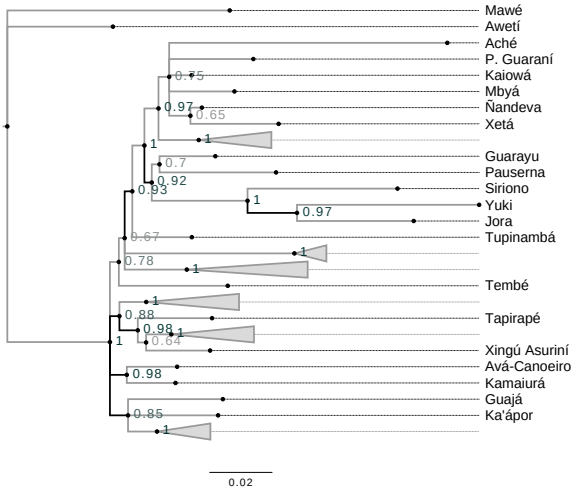

Fig. 2. Tupí-Guaraní: Majority-rule consensus tree of binary root-meaning coding

both sets of results have not been previously suggested in the literature, making an evaluation of which coding method brings us closer to the 'true' tree difficult. In any case, the coding method chosen has serious implications for the classification.

B. Chapacuran dataset

For the Chapacuran dataset, no significant difference is observed between the binary root-meaning coding and the binary cognate coding of the dataset. In both resulting trees all clades are well supported with over 0.9 posterior probability. At the same time, both coding procedures produce results that conform to the comparative method classification in [8], while they add more internal structure to the tree. For the sake of space, only the maximum clade credibility (MCC) tree of the analysis of the binary root-meaning coding is presented here in Figure 3.

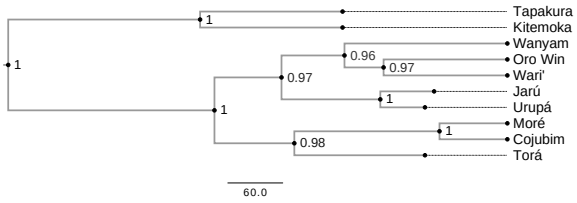

Fig. 3. Chapacuran: MCC tree of binary root-meaning coding

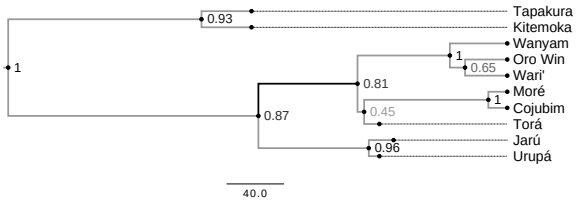

Fig. 4. Chapacuran: MCC tree of multistate meaning-based coding

In comparison with the two previously discussed coding procedures, the multistate meaning-based coding produces the most divergent results. Most surprisingly, the Waric branch becomes paraphyletic through the insertion of the Moreic branch (Moré-Cojubim-Torá) between the Oro Win-Wari'-Wanyam and Urupá-Jarú clades. This configuration does not agree with the comparative method classification. Additionally, the multistate coding method fails to recover the Moreic branch and the Wari'-Oro Win clade (they are present in the tree with very low posterior probabilities). Both of these clades are well supported in the other analyses, and the Moreic branch is well established by the comparative method as well.

V. Discussion

The different coding methods produced significant differences in topology and posterior probabilities for both datasets. However, no obvious trends have emerged at this point regarding convergence in results of the coding methods. For the Tupí-Guaraní dataset, the multistate meaning-based and binary root-meaning coding produced the most similar results topologically, while for the Chapacuran dataset the binary root-meaning and the cognate coding results were the most similar.

According to [11], the main differences between multistate meaning-based and binary root-meaning codings are higher support values and shorter branch lengths for the latter. Indeed, for Tupí-Guaraní the binary root-meaning coding produced much higher posterior probabilities than the multistate meaning-based coding, with branch lengths reduced by a factor of 10 (data not shown). This is also largely true for the Chapacuran dataset. However, while the multistate tree contains clades with much lower posterior probability than the binary root-meaning tree, there is a large number of clades in both trees with high posterior probabilities. No direct comparison can be made for the branch lengths, since the Chapacuran trees are time-trees.

For both datasets, the prediction of [11] that no topological differences are expected due to the binary recoding of

multistate characters does not hold. For the Tupí-Guaraní dataset, the two trees were compatible, meaning that there was no highly supported node in the binary root-meaning tree that was contradicted in the multistate meaning-based tree. However, the very low posterior probabilities almost throughout the multistate tree mean that many nodes that were highly supported in the binary root-meaning tree were completely absent from the multistate tree, leading to significant topological differences. For the Chapacuran dataset, the topological differences were even more striking: both trees contained highly supported incompatible clades, leading to directly contradicting topologies. At the same time, two highly supported clades in the binary root-meaning tree are not supported in the multistate meaning-based tree.

Regarding the comparison with the results of the binary cognate coding, again no clear trends can be identified. For Chapacuran the differences between binary root-meaning coding and binary cognate coding are negligible, while for Tupí-Guaraní these two coding methods produce the most divergent results with clearly contradicting topologies.

There are various potential reasons why the two datasets behave differently that need to be investigated in detail in order to be fully understood. The two datasets used vary greatly in size (both regarding the number of languages and the number of meanings), in the amount of identified semantic shift (much higher in the Tupí-Guaraní dataset), and in the number of synonyms included for the same meaning (high in Tupí-Guaraní, almost non-existent in Chapacuran). Furthermore, different coding methods may be affected to various degrees by linguistic phenomena, such as parallel semantic shifts (binary root-meaning and multistate meaning-based coding), and methodological artifacts, such as spurious subgroups supported by shared 'absences' (binary root-meaning coding).

VI. Conclusion

Our empirical test of three different coding methods on two lexical datasets showed that coding method choice can have a significant (and unpredictable) impact on the resulting topology and posterior probabilities, contrary to prior expectations [11]. At the same time, the exact reasons and mechanisms underlying the observed differences in the results are not fully understood and require further investigation.

Acknowledgments

We thank our colleagues past and present in the Tupí-Guaraní Comparative Project: Vivian Wauters, Keith Bartolomei, Erin Donnelly, and Michael Roberts. We are indebted to those who have shared unpublished data with us: Sebastian Drude, Noé Gasparini, Françoise Rose, Eva-Maria Rößler, and Rosa Vallejos. We benefited from conversations with Paul Heggarty and participants in the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics. Diamantis Sellis facilitated the automated binary coding of the dataset and developed scripts to verify consistency between comparative and cognate lists.

References

[1] P. J. Lyon, Ed., *Native South Americans: Ethnology of the Least Known Continent*. Prospect Heights: Waveland Press, 1974.

[2] P. Epps, "Language Classification, Language Contact, and Amazonian Prehistory," *Language and Linguistics Compass*, vol. 3, no. 2, pp. 581–606, 2009.

[3] C. Everett, "A Survey of Contemporary Research on Amazonian Languages," *Language and Linguistics Compass*, vol. 4, no. 5, pp. 319–336, 2010.

[4] L. Campbell, "Classification of the indigenous languages of South America," in *The Indigenous Languages of South America: A comprehensive guide*, L. Campbell and V. Grondona, Eds. Berlin: De Gruyter Mouton, 2012, pp. 59–166.

[5] H. Hammarström, "Basic vocabulary comparison in South American languages," in *The Native Languages of South America: Origins, Development, Typology*, L. O'Connor and P. Muysken, Eds. Cambridge: Cambridge University Press, 2014, pp. 56–72.

[6] R. S. Walker and L. A. Ribeiro, "Bayesian phylogeography of the Arawak expansion in lowland South America," *Proceedings of the Royal Society of London, Series B*, vol. 278, no. 1718, pp. 2562–2567, 2011.

[7] L. Michael, N. Chousou-Polydouri, K. Bartolomei, E. Donnelly, V. Wauters, S. Meira, and Z. O'Hagan, "A Bayesian Phylogenetic Classification of Tupí-Guaraní," *LIAMES*, vol. 15, no. 2, pp. 193–221, 2015.

[8] J. Birchall, M. Dunn, and S. J. Greenhill, "A combined comparative and phylogenetic analysis of the Chapacuran language family," *International Journal of American Linguistics*, p. to appear, 2016.

[9] W. Hennig, *Phylogenetic systematics*. Urbana: University of Illinois Press, 1966.

[10] K. Rexová, D. Frynta, and J. Zrzavý, "Cladistic analysis of languages: Indo-European classification based on lexicostatistical data," *Cladistics*, vol. 19, no. 2, pp. 120–127, Apr. 2003.

[11] M. Pagel and A. Meade, "Estimating rates of lexical replacement on phylogenetic trees of languages," in *Phylogenetic methods and the prehistory of languages*, P. Forster and C. Renfrew, Eds. Cambridge: McDonald Institute for Archaeological Research, 2006, pp. 173–182.

[12] W. Chang, C. Cathcart, D. Hall, and A. Garrett, "Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis," *Language*, vol. 91, no. 1, pp. 194–244, 2015.

[13] R. D. Gray and Q. D. Atkinson, "Language-tree divergence times support the Anatolian theory of Indo-European origin," *Nature*, vol. 426, pp. 435–439, 2003.

[14] C. Bowern and Q. D. Atkinson, "Computational phylogenetics and the internal structure of Pama-Nyungan," *Language*, vol. 88, no. 4, pp. 817–845, 2012.

[15] J. Felsenstein, *Inferring Phylogenies*. Sunderland, MA: Sinauer Associates, Inc, 2004.

[16] P. O. Lewis, "A likelihood approach to estimating phylogeny from discrete morphological character data," *Systematic Biology*, vol. 50, no. 6, pp. 913–925, 2001.

[17] J. Huelsenbeck and F. Ronquist, "MRBAYES: Bayesian inference of phylogenetic trees," *Bioinformatics*, no. 8, pp. 754–755.

[18] F. Ronquist and J. Huelsenbeck, "MrBayes 3: Bayesian phylogenetic inference under mixed models," *Bioinformatics*, 2003.

[19] A. D. Rodrigues and W. Dietrich, "On the linguistic relationship between Mawé and Tupí-Guaraní," *Diachronica*, vol. 14, no. 2, pp. 265–304, 1997.

[20] S. Meira and S. Drude, "A summary reconstruction of Proto-Maweti-Guaraní segmental phonology," *Boletim do Museu Paraense Emílio Goeldi, Ciências Humanas*, vol. 10, no. 2, pp. 275–296, 2015.

[21] A. J. Drummond, M. A. Suchard, D. Xie, and A. Rambaut, "Bayesian phylogenetics with BEAUti and the BEAST 1.7," *Molecular Biology and Evolution*, vol. 29, no. 8, pp. 1969–1973, 2012.

[22] A. D. Rodrigues and A. S. A. C. Cabral, "Revendo a classificação interna da família tupi-guarani," in *Línguas Indígenas Brasileiras. Fonologia, Gramática e História. Atas do I Encontro Internacional do GTLI*, A. S. A. C. Cabral and A. D. Rodrigues, Eds. Belém: Editora Universitaria da UFPA, 2001, vol. I, pp. 327–337.