Phylogenetic inference of reticulate evolution in Tupí-Guaraní languages

Tupí-Guaraní is the largest branch of Tupían, one of the largest linguistic families of the Americas [1]. Classifications for establishing its internal relations and chronology of splits vary significantly in their proposed trees [2-9], frequently in disagreement with the history suggested by archeology and material culture [10-12]. The hypothesis that the Tupí-Guaraní family is a reticulate set of languages has been argued since the earliest classification studies [3-9], and is supported by a growing body of comparative linguistic data, historical evidence, and the possibility of multiple events of hybridization in the past. Such events, far more intensive than mere lexical borrowings, are likely the cause for long-standing uncertainties in the position of languages like Avá-Canoeiro, Kamajurá, and Tenetehara. In this work, we use phylogenetic methods to test the hypothesis of reticulate evolution and to infer the phylogeny of the family.

While modern computational approaches address the limits imposed by reticulated events in linguistics (e.g. [13-14]), sometimes making them their focus (e.g., [15-17]), most classifications maintain the tree model as their goal, in a practice that has led to an interesting theoretical debate in recent years [18-20], influenced by the rising prominence of networks in biological phylogenetics. We argue that, once the major traits of vertical inheritance have been detected, networks can be more suitable for the study of reticulate phylogenies, because they better reflect the evolutionary process of hybridization, which is not well captured by the available tree-based methods. The reconstruction of such evolution requires adopting a set of practices and tools that have been developed for biological classification but are still under-exploited in historical linguistics.

We present the results of an ongoing research, in which we develop on and extend a previous study of ours (under review) that inferred a phylogenetic tree for Tupí-Guaraní using data from [21] and following common practices in computational historical linguistics [22-24]. Our analysis focuses on the relation of the Southern branches ("Guaraní", the last split in almost all linguistic classifications, but associated with the oldest ceramics) and the position of Tupinambá in the tree, as well as the detection of signals compatible with events of past hybridization. We reanalyze our data with phylogenetic comparative methods for reticulations [25], especially methods for inferring phylogenetic networks with maximum pseudo-likelihood under incomplete lineage sorting [26] and bipartition dissimilarity [27]. The results are compared with our previous findings and discussed from an expert perspective on the history of the family, supported by an automatic identification of which characters are involved in the inferred events of hybridization (including borrowings).

Our results support the hypothesis of a partially reticulate evolution and shed light on the internal history of Tupí-Guaraní, facilitating the integration of linguistic and archeological evidence and laying foundations for similar research with other families of the same geographic area. Whether the network is an appropriate representation of the evolution of Tupí-Guaraní remains an open question, including how much it might clarify disagreements with archeology. It is our opinion, however, that this is one of the first steps towards a better integration of linguistic and archeological data that can be used to understand the hybridization process and the nature of the languages involved.

References

[1] Rodrigues AD, Cabral A. Tupían. In: Campbell L, Grondona V, (ed.). The Indigenous Languages of South America. De Gruyter, Berlin/Boston; 2012. p. 495-574.
[2] Galucio AV et al. Genealogical relations and lexical distances within the Tupian linguistic family. Boletim do Museu Paraense Emílio Goeldi Ciências Humanas. 2015; 10(2):229-274.
[3] Rodrigues AD. Relações internas na família lingüística Tupí-Guaraní. Revista de antropologia. 1984; 27/28:33-53.
[4] Dietrich W. More evidence for an internal classification of Tupí-Guaraní languages. Berlin: Gebr. Mann; 1990.
[5] Rodrigues AD, Cabral ASAC. Revendo a classificação interna da família Tupí-Guaraní. Línguas Indígenas Brasileiras. Atas do I Encontro Internacional do GTLI da ANPOLL. 2002.
[6] Mello AAS. Estudo histórico da família linguística Tupí-Guaraní: Aspectos fonológicos e lexicais [PhD thesis]. Universidade Federal de Santa Catarina.Florianópolis; 2000.
[7] Mello AAS. Evidências fonológicas e lexicais para o sub-agrupamento interno Tupí-Guaraní; 2002. Línguas Indígenas Brasileiras.Atas do I Encontro Internacional do GTLI da ANPOLL.
[8] Michael LD et al. A Bayesian phylogenetic classification of Tupí-Guaraní. LIAMES. 2015; 15(2):193-221.
[9] Gerardi F, Reichert S. The Tupí-Guaraní language family: A phylogenetic classification. Diachronica. 2021; 38(2):151-188.
[10] Noelli FS. The Tupí: Explaining origin and expansions in terms of archaeology and of historical linguistics. Antiquity. 1998; 72(277):648-663.
[11] Brochado JP. An ecological model of the spread of pottery and agriculture into Eastern South America [PhD thesis]. University of Illinois at Urbana-Champaign; 1984.
[12] Noelli FS. A ocupação humana na Região Sul do Brasil: Arqueologia, debates e perspectivas-1872-2000. Revista USP. 1999;44:218-269.
[13] Greenhill et al. Does horizontal transmission invalidate cultural phylogenies? Proceedings of the Royal Society B: Biological Sciences 276.1665 (2009): 2299-2306.
[14] Bowern C et al. Does lateral transmission obscure inheritance in hunter-gatherer languages?. PloS one, 2011, 6(9), e25195.
[15] Verkerk A. Detecting non-tree-like signal using multiple tree topologies. Journal of Historical Linguistics, 2019, 9(1), 9-69.
[16] Nelson-Sathi S et al. Networks uncover hidden lexical borrowing in Indo-European language evolution. Proceedings of the Royal Society B. 2011; 278(1713), 1794-1803.
[17] List JM et al. Using phylogenetic networks to model Chinese dialect history. Language Dynamics and Change, 2014, 4(2), 222-252
[18] Jacques G, List JM. Save the trees: Why we need tree models in linguistic reconstruction (and when we should apply them). Journal of historical linguistics, 2019, 9(1), 128-167.
[19] Kalyan S, François A. When the waves meet the trees: A response to Jacques and List. Journal of Historical Linguistics, 2019, 9(1), 168-177.

[20] Evans CL et al. The uses and abuses of tree thinking in cultural evolution. Philosophical Transactions of the Royal Society B, 2021, 376(1828).
[21] Gerardi F et al. TuLeD: Tupían lexical database; 2021. Available from: https://tular.clld.org/contributions/tuled
[22] List JM et al. Sequence comparison in computational historical linguistics. Journal of Language Evolution, 2018, 3(2), 130-144.
[23] Jäger G. Computational historical linguistics. Theoretical Linguistics, 2019, 45(3-4), 151-182.
[24] Hoffmann K et al. Bayesian phylogenetic analysis of linguistic data using BEAST. Journal of Language Evolution, 2021, 6(2), 119-135.
[25] Bastide P et al. Phylogenetic Comparative Methods for Phylogenetic Networks with Reticulations. Systematic Biology, 2018, 67(5):800-820.
[26] Solís-Lemus C, Ané C. Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting. PLoS Genet, 2016, 12(3).
[27] Boc A et al.Inferring and validating horizontal gene transfer events using bipartition dissimilarity, Systematic Biology, 2010, 59: 195-211.