A Phylogenetic Study of the Cariban Family: Combining Linguistic and Archaeological Data

This study presents a preliminary linguistic phylogenetic analysis of the Cariban language family, a group of indigenous languages in Northern South America and Central Brazil. The family comprises approximately three dozen languages, most of which are spoken by small communities of a few hundred speakers (Hammarström et al., 2022). The languages are closely related in phonology and lexicon. However, except for a few shallow clades, most of the family's history is unknown and subject to many theories on its homeland and expansion routes (Meira & Franchetto, 2005; Gildea, 2012).

We follow the best practices in current computational historical linguistics (Hoffman et al., 2021; Greenhill et al., 2020; Jäger, 2019; Tresoldi et al., 2022), including initial analyses with neighbournets (Huson & Bryant, 2006) and Bayesian MCMC inference (Bouckaert et al., 2019) using different evolutionary models. We collected the data from reliable sources (Matter, 2020; de Tauste, 1680; Courtz, 2008; Largo, 2011; Ruiz Blanco 1888 [1690], von den Steinen 1892; Koehn & Koehn, 1986). It is organised into an independent, normalised, and open-access database in CLDF format (Forkel et al., 2018), carrying cognate assignments made by experts (Carvalho et al., forth.). We will discuss the results of our classification in the context of earlier classifications of the Cariban language family, including those by Derbyshire (1999), Meira (2006), Gildea (2012), and Meira et al. (2015), such as the statistical support of our findings for consensual and nearly consensual clades (e.g., Parukotoan, Pekodian) and for other clades proposed in the literature (e.g., the "Venezuelan" branch, Wayana-Apalai, Panare-Pemongan). Besides our maximum clade credibility (MCC) tree informed by linguistic and archaeological data, we will present our first phylogeographic inference models (Lemey et al., 2009).

This study is part of a larger initiative to analyse the linguistic history of South America. It will contribute analyses and insights into the evolution and relationships within the Cariban language family, including the location of its homeland, the date of its first expansions, and its migration movements. For general phylogenetic inference, it will contribute to the discussion on the solution and strategies for incorporating archaeological data, a necessary step when classifying language families without written records extending over multiple centuries. The classification based on lexical evidence will motivate the search for shared innovations in phonology and morphology, paving the way for the reconstruction of intermediate-level proto-languages using traditional and computer-assisted methods. It will likewise expand the foundations for research on language contact among South American native languages.

References

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., et al. (2019) "BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis". *PLoS computational biology*, 15(4), e1006650. https://doi.org/10.1371/journal.pcbi.1006650

Courtz, H. (2008). *A Carib grammar and dictionary*. Magoria books.

de Tauste, F. (1680). *Arte y bocabulario de la lengua de los indios chaymas, cumanagotos, cores, parias, y otros diversos de la provincia de Cumana, o Nueva Andalucia*. Madrid: Bernardo de Villa.

Derbyshire, D. (1999). "Carib". In Dixon, R. M. W. & Aikhenvald, A. (eds.), *The Amazonian Languages*, 23-64. Cambridge: Cambridge University Press.

Carvalho, F. O., Ferraz Gerardi, F., Aragon, C. & Tresoldi, T. (forth.). *Cariban Lexical Database*.

Forkel, R., List, J.-M., Greenhill, S. J., Bank, S., Rzymski, C., Cysouw, M., Hammarström, H., Haspelmath, M., Kaiping, G. A. & Gray, R. D. (2018). "Cross-linguistic Data Formats, advancing data sharing and reuse in comparative linguistics". *Scientific Data*, 5:180205. https://doi.org/10.1038/sdata.2018.205

Gildea, S. (2012). "Linguistic studies in the Cariban family". In Campbell, L. & Grondona, V. (eds.), *The Indigenous Languages of South America: A Comprehensive Guide*, 441-494. Berlin, Boston: Berlin: Mouton.

Hammarström, H., Forkel, R., Haspelmath, Martin & Bank, S. (2022). *Glottolog 4.7*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.7398962 . Available online at http://glottolog.org, accessed on January 17th, 2023.

Hoffmann, K., Bouckaert, R., Greenhill, S. J. & Kühnert, D. (2021). "Bayesian phylogenetic analysis of linguistic data using BEAST". *Journal of Language Evolution*, 6: 119–135. https://doi.org/10.1093/jole/lzab005

Huson, D. H. & Bryant, D. (2006). "Application of Phylogenetic Networks in Evolutionary Studies". *Molecular Biology and Evolution*, 23(2): 254-267. https://doi.org/10.1093/molbev/msj030

Greenhill, S. J., Heggarty, P., & Gray, R. D. (2020). "Bayesian Phylolinguistics". In Janda, R. D., Joseph, B. D. & Vance, B. S. (eds) *The Handbook of Historical Linguistics*, Volume II, pp. 226–253. Wiley-Blackwell: New Jersey.

Jäger, G. (2019). “Computational historical linguistics”. *Theoretical Linguistics*, 45(3-4): 151-182. https://doi.org/10.1515/tl-2019-0011

Koehn, E., & Koehn, S. (1986). “Apalai”. In D. S. Derbyshire & G. K. V. I. Pullum (Eds.), *Handbook of Amazonian languages*, Vol. 17: 33-127. Berlin: de Gruyter.

Largo, W. (2011). *Una Gramática del Yukpa (Colombia)*. Privately published.

Lemey, P., Rambaut, A., Drummond, A. J. & Suchard, M. A. (2009). “Bayesian Phylogeography Finds Its Roots”. *PLoS Computational Biology* 5, e1000520. https://doi.org/10.1371/journal.pcbi.1000520

Matter, F. (2020). *Comparative Cariban Database. Leipzig: Max Planck Institute for Evolutionary Anthropology*. Available online at https://cariban.clld.org/, accessed on January 17th, 2023.

Meira, S. (2006). “Cariban Languages”. In Brown, K. (ed.), *Encyclopedia of Language and Linguistics*, 199-203. 2nd edn. Amsterdam: Elsevier.

Meira, S. and B. Franchetto (2005). The Southern Cariban languages and the Cariban family. *International Journal of American Linguistics* 71: 127-192.

Meira, S., Birchall, J., Chousou-Polydouri, N. (2015). “A character-based internal classification of the Cariban family”. Talk. *48th Annual Meeting of the Societas Linguisticae Europaea*. Leiden, Netherlands. Available online at https://www.academia.edu/15980095/A_character-based_internal_classification_of_the_Cariban_language_family, accessed on January 17th, 2023.

Ruiz Blanco, M. (1888 [1690]). *Arte y tesoro de la lengua cumanagota (Vol. 3)*. Leipzig, Germany: B.G. Teubner.

von den Steinen, K. (1892). *Die Bakaïrí-Sprache: Wörterverzeichnis, Sätze, Sagen, Grammatik mit beiträgen zu einer Lautlehre der Karaïbischen Grundsprache*. Leipzig: K. F. Koehler's Antiquarium.

Tresoldi, T., Rzymski, C., Forkel, R., Greenhill, S. J., List, J.-M. & Gray R. (2022). “Managing historical linguistic data for computational phylogenetics and computer-assisted language comparison”. In Berez-Kroeker, A. L., McDonnell, B., Koller, E. & Collister, L. B. (eds). *Open Handbook of Linguistic Data Management*. MIT Press. https://doi.org/10.7551/mitpress/12200.003.0033