Classifying Tupí-Guaraní Languages

Fabrício F. Gerardi and Stanislav Reichert

Abstract

According to Ethnologue [9], there are 456 languages spoken in South America. Spanning Brazil, Paraguay, Bolivia, French Guiana, Peru and counting more than 5 million speakers, Tupían languages are by far the most widely spoken languages on this continent in terms of territory [13]. Tupí-Guaraní (TG) is the largest Tupían subfamily with ca. 48 languages. While Paraguaian Guarani has ca. 5 million speakers, many of these languages, according to the most recent Brazilian census in 2010 [7], have less than 100 speakers and a dozen of these have probably died out since then. Tupí-Guarani languages are also territorially widespread like no other language family in South America. These languages seem to have originated ca. 3000 years B.P., possibly in what is today the Brazilian state of Rondonia. As of today, the exact place of their origin and the routes of their spread through their historical territories remain topic of an ongoing scientific debate [12, 15].

The last ten years have seen a significant increase in the amount of studies of Tupí-Guaraní languages. Nonetheless comprehensive descriptions are still at large for many of these languages, which share very interesting and sometimes rare typological and grammatical characteristics. These characteristics are of importance, since they allow us not only to make inferences about the evolution of these languages and to refine our knowledge of linguistic typology, but are also useful for gaining knowledge about linguistic areas in South America [5]. They are also relevant as they furnish data that allow for the classification of languages using quantitative methods in a more reliable way.

Comparative linguistic work may throw some light on contact and migrations, thus emphasizing or showing disagreement with the interpretation of archaeological data [14, 12]. The lack of a consensus regarding the TG homeland and its dispersal in both archeology and linguistics [12, 1, 11] requires this kind of research. It is, however, clear that the answer to the questions of how, where and when the dispersion occurred will eventually come from combined insights from both disciplines [15, 1, 11], as it has been the case for other language families [2, 10].

We have collected 280 words from 39 Tupí-Guaraní languages and 2 additional Tupían languages to be used as outliers in order to classify Tupí-Guaraní languages lexically and phonologically paying special attention to the lowest levels of the tree (subgrouping). In order to classify the lexicon, the first stage of our work applies Bayesian phylogenetic methods taken from molecular biology. We also use computational methods commonly employed in historical linguistics to produce an alignment

of words in the various languages in order to calculate lexical distances according to a recent proposed method [8].

Another part of our work comprises a separate attempt to classify these languages based on grammatical properties, updating and extending (more languages and more grammatical properties) a paper by Wolf Dietrich [6] and finally combining these results with those of a double lexical analysis. Our research also tries to combine archaeological and recent data from genetic studies with linguistic data in order to allow for inferences on a more solid ground for their classification and spread.

Finally we will show that the data allows us to conjecture the about geographical dispersion of the TG groups departing from the so called "linguistic migration" [4] and showing how the data collected and the results of our analysis can be used to challenge or confirm archaeological proposals regarding movements or migrations [3] of these populations.

References

[1] F. O. d. Almeida and E. G. Neves. Evidências arqueológicas para a origem dos tupi-guarani no leste da amazônia. *Mana*, 21(3):499–525, 2015.

[2] D. W. Anthony. *The horse, the wheel, and language: how Bronze-Age riders from the Eurasian steppes shaped the modern world*. Princeton University Press, 2010.

[3] P. Bellwood. *First migrants: ancient migration in global perspective*. John Wiley & Sons, 2014.

[4] L. Campbell. *Historical linguistics*. Edinburgh University Press, 2013.

[5] M. Crevels and H. Van der Voort. The guaporé-mamoré region as a linguistic area. In P. muysken, editor, *From linguistic areas to areal linguistics*, pages 151–179. John Benjamins, Amsterdam, 2008.

[6] W. Dietrich. *More evidence for an internal classification of tupi-guarani languages*. Gebr. Mann Verlag, 1990.

[7] IBGE. Os indígenas no censo demográfico 2010. `https://indigenas.ibge.gov.br/images/pdf/indigenas/folder_indigenas_web.pdf`. Accessed: 2017-27-12.

[8] G. Jäger. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. *Language Dynamics and Change*, 3(2):245–291, 2013.

[9] M. P. Lewis, G. F. Simons, C. D. Fennig, et al. *Ethnologue: Languages of the world*, volume 16. SIL international Dallas, TX, 2009.

[10] J. P. Mallory. *The origins of the Irish*. Thames & Hudson, 2013.

[11] A. A. S. Mello and A. Kneip. Novas evidências linguísticas (e algumas arqueológicas) que apontam para a origem dos povos tupi-guarani no leste amazônico. *Literatura y lingüística*, (36):299–312, 2017.

[12] F. S. Noelli. The tupi expansion. In H. Silverman and W. Isbell, editors, *The Handbook of South American Archaeology*, pages 659–670. Springer, 2008.

[13] A. Rodrigues. Tupían. In L. Campbell and V. Grondona, editors, *The indigenous languages of South America: a comprehensive guide*, volume 2. Walter de Gruyter, 2012.

[14] A. D. Rodrigues. Hipótese sobre as migrações dos três subconjuntos meridionais da família tupi-guarani. In *Atas do II Congresso Nacional da ABRALIN*, pages 1596–1605, 2000.

[15] G. Urban. On the geographical origins and dispersion of tupian languages. *Revista de Antropologia*, pages 61–104, 1996.