Incorporating typological features in tree inference: Lessons from phonotactics

Jayden Macklin-Cordes
Laboratoire Dynamique Du Langage, Centre National de la Recherche Scientifique, Lyon, France.

Keywords: *Linguistic phylogenetics, quantitative methods, historical linguistics, phonology, Pama-Nyungan*

There is a veritable cornucopia of facts and features of human languages that can be encoded in comparative datasets. However, inferring genealogical trees of languages from this data is challenging. In this paper, I discuss how (not) to incorporate typological data into the task of tree inference, using a wayward example of phonotactics.

Computationally inferred language phylogenies make up a critical component of an expanding range of studies on human language and human history. Beyond genealogical classification of languages, language phylogenies make up an essential component of an increasing range of studies into linguistic typology (e.g. Jäger & Wahle 2021), evolutionary dynamics of human language (e.g. Allassonnière-Tang & Dunn 2020) and ancient human migrations (e.g. Robbeets et al. 2021).

One current limitation is an almost exclusive reliance on lexical cognate data to infer language phylogenies. Ideally, language phylogenies would be inferred from a more encompassing range of data including phonology and grammar. Previous studies have identified *phylogenetic signal*, a correlation between variance in the data and expected variance given an independent predefined reference phylogeny, in frequency data from biphones (two-segment sequences), a basic kind of representation of phonotactics (Dockum 2018; Macklin-Cordes, Bowern & Round 2021). This finding motivated the hypothesis of the current study, that this phonotactic data could be a useful extra source of information from which to infer language trees, in combination with existing cognate data.

Here, I present the results of a proof-of-concept assessment of Bayesian phylogenetic tree inference using biphone frequency data in concert with cognate data, tested on 44 western Pama-Nyungan languages. This 'combined' model is evaluated against a 'separate' model in which trees are inferred from each data type separately, and a 'cognates only' model (essentially replicating a subset of the Pama-Nyungan phylogeny inferred by Bouckaert, Bowern & Atkinson 2018). Far from strengthening tree inference, the addition of phonotactics causes both 'combined' and 'separate' models to become unstable and both fail to converge around a consistent result. Further, qualitative comparison of the trees produced by the 'combined' and 'cognates only' models shows that the addition of phonotactics reduces informative branching structure across the tree and moving a language with a particularly distinct phonotactic profile (Western Arrernte) to a distant outgroup position in the tree.

I conclude with a two-part discussion. Firstly, I address the technical limitations of the current implementation and present options for future evaluation of quantitative phonotactics. This will require addressing the assumption that all biphone variables operate independently by identifying patterns of interdependence within the data, using methods such as phylogenetic principal components analysis (pPCA) (Revell 2009) and phylogenetic factor analysis (Hassler et al. 2021). This is a hot topic in evolutionary biology as well as linguistics and subject to active methodological development. Secondly, I outline a more generalisable set of steps for proof-of-concept evaluation of other novel data sources for phylogenetic tree inference. A key task here is linking stationary frequency distributions to diachronic processes (along the lines discussed in Macklin-Cordes & Round 2020) and building these directly into evolutionary models used by Bayesian phylogenetic algorithms.

References

Allassonnière-Tang, Marc & Michael Dunn. 2020. The evolutionary trends of grammatical gender in Indo-Aryan languages. *Language Dynamics and Change* 11(2). 211–240. https://doi.org/10.1163/22105832-bja10011.

Bouckaert, Remco R., Claire Bowern & Quentin D. Atkinson. 2018. The origin and expansion of Pama–Nyungan languages across Australia. *Nature Ecology & Evolution* 2(4). 741--749. https://doi.org/10.1038/s41559-018-0489-3.

Dockum, Rikker. 2018. Phylogeny in phonology: How Tai sound systems encode their past. *Proceedings of the Annual Meetings on Phonology* 5. https://journals.linguisticsociety.org/proceedings/index.php/amphonology/article/view/4238 (21 February, 2018).

Hassler, Gabriel W., Brigida Gallone, Leandro Aristide, William L. Allen, Max R. Tolkoff, Andrew J. Holbrook, Guy Baele, Philippe Lemey & Marc A. Suchard. 2021. Principled, practical, flexible, fast: a new approach to phylogenetic factor analysis. *arXiv:2107.01246 [q-bio, stat]*. http://arxiv.org/abs/2107.01246 (22 December, 2021).

Jäger, Gerhard & Johannes Wahle. 2021. Phylogenetic Typology. *Frontiers in Psychology* 12. 2852. https://doi.org/10.3389/fpsyg.2021.682132.

Macklin-Cordes, Jayden L., Claire Bowern & Erich R. Round. 2021. Phylogenetic signal in phonotactics. *Diachronica*. John Benjamins 38(2). 210–258. https://doi.org/10.1075/dia.20004.mac.

Macklin-Cordes, Jayden L. & Erich R. Round. 2020. Re-evaluating Phoneme Frequencies. *Frontiers in Psychology*. Frontiers 11. https://doi.org/10.3389/fpsyg.2020.570895.

Revell, Liam J. 2009. Size-Correction and Principal Components for Interspecific Comparative Studies. *Evolution* 63(12). 3258–3268. https://doi.org/10.1111/j.1558-5646.2009.00804.x.

Robbeets, Martine, Remco Bouckaert, Matthew Conte, Alexander Savelyev, Tao Li, Deog-Im An, Ken-ichi Shinoda, et al. 2021. Triangulation supports agricultural spread of the Transeurasian languages. *Nature* 1–6. https://doi.org/10.1038/s41586-021-04108-8.