4 Phlorest: A Database of Consistent and Reusable Language Phylogenies

Robert Forkel[1], Simon Greenhill[2]

[1] Dep. of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology Leipzig
[2] School of Psychology, University of Auckland, Auckland

The last few decades have seen the publication of many language phylogenies. These phylogenies have proven to be incredibly powerful tools for making inferences about language relationships (e.g. Gray, Drummond, and Greenhill 2009; Kolipakam et al. 2018; Remco R. Bouckaert, Bowern, and Atkinson 2018; Chang et al. 2015; Greenhill et al. 2022), or as a backbone for testing hypotheses about language change (e.g. Dunn et al. 2011), linguistic reconstructions (e.g. Carling and Cathcart 2021), and evolutionary processes (e.g. Greenhill et al. 2017). Often the results of these phylogenetic studies are repurposed by other researchers to test other hypotheses Watts et al. (2016). Or the results themselves are controversial e.g. witness the arguments about the age of Indo-European Chang et al. (2015) or the debates about language universals Dryer (2011).

We therefore need good ways for researchers to obtain, inspect, compare them, and reuse these phylogenies. However, to date this re-use is hard, often requiring detailed phylogenetic knowledge to identify the relevant files, understand their formats, and extract the critical information. Phlorest is a database of published language phylogenies that aims to standardise the outputs of these analyses to make them Findable, Accessible, Interoperable, and Reusable (Wilkinson et al. 2016). Phlorest collects published language phylogenies into a single database in a consistent and easily usable format (CLDF, Forkel et al. 2018). Currently, Phlorest contains 42 phylogenies, covering a total of 4266 varieties from 2172 languages. Each analysis is preprocessed to a consistent format, providing a summary tree and a posterior tree sample, linked where possible to the raw data. Each taxon in the analysis is mapped to catalogues like Glottolog (https://glottolog.org) and D-PLACE (https://d-place.org/) so that users can readily identify which languages were included in each analysis.
In this talk we will present Phlorest and discuss the benefits it provides. First, phlorest enables replicability and reuse of these trees. Second, having these phylogenies aligned in time and space enables us to compare patterns and processes across the globe. Third, phlorest allows us to scale up to bigger questions by combining trees into super trees. Finally, phlorest allows us to highlight interesting big picture findings from historical linguistics to the wider public, providing a highly visible resource that brings this research to a wider audience.

Bouckaert, Remco R., Claire Bowern, and Quentin D. Atkinson. 2018. “The Origin and Expansion of Pama-Nyungan Languages Across Australia.” Nature Ecology & Evolution. https://doi.org/10.1038/s41559-018-0489-3.

Bouckaert, Remco R, et al.. 2012. “Mapping the Origins and Expansion of the Indo-European Language Family.” Science 337 (6097): 957–60. https://doi.org/10.1126/science.1219669.

Carling, Gerd, and Chundra Cathcart. 2021. “Reconstructing the Evolution of Indo-European Grammar.” Language 97 (3): 561–98. https://doi.org/10.1353/lan.2021.0047.

Chang, Will, Chundra Cathcart, David Hall, and Andrew Garrett. 2015. “Ancestry-Constrained Phylogenetic Analysis Supports the Indo-European Steppe Hypothesis.” Language 91 (1): 194–244. https://doi.org/10.1353/lan.2015.0005.

Dryer, Matthew S. 2011. “The Evidence for Word Order Correlations.” Linguistic Typology 15: 335–80. https://doi.org/10.1515/LITY.2011.024.

Dunn, Michael, Simon J. Greenhill, S. C. Levinson, and Russell D. Gray. 2011. “Evolved Structure of Language Shows Lineage-Specific Trends in Word-Order Universals.” Nature 473 (7345): 79–82. https://doi.org/10.1038/nature09923.

Forkel, Robert, et al. 2018. “Cross-Linguistic Data Formats, Advancing Data Sharing and Re-Use in Comparative Linguistics.” Scientific Data 5 (1): 180205. https://doi.org/10.1038/sdata.2018.205.

Gray, Russell D., Alexei J Drummond, and Simon J. Greenhill. 2009. “Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement.” Science 323 (5913): 479–83. https://doi.org/10.1126/science.1166858.

Greenhill, Simon J., Hannah J. Haynie, Robert M Ross, Angela M. Chira, Johann-Mattis List, Lyle Campbell, Carlos A. Botero, and Russell Gray. 2022. “A Recent Northern Origin for the Uto-Aztecan Family,” August. https://doi.org/10.31235/osf.io/k598j.

Greenhill, Simon J., Chieh-Hsi Wu, Xia Hua, Michael Dunn, Stephen C. Levinson, and Russell D. Gray. 2017. “Evolutionary Dynamics of Language Systems.” Proceedings of the National Academy of Sciences, 201700388. https://doi.org/10.1073/pnas.1700388114.

Kolipakam, Vishnupriya, Fiona M Jordan, Michael Dunn, Simon J Greenhill, Remco Bouckaert, Russell D Gray, and Annemarie Verkerk. 2018. “A Bayesian Phylogenetic Study of the Dravidian Language Family.” Royal Society Open Science 5: 171504. https://doi.org/10.1098/rsos.171504.

Levy, Roger, and H Daumé III. 2011. “Computational Methods Are Invaluable for Typology, but the Models Must Match the Questions.” Linguistic Typology 15: 393–99. https://doi.org/10.1515/LITY.2011.026.

Watts, Joseph, Oliver Sheehan, Quentin D. Atkinson, Joseph Bulbulia, and Russell D. Gray. 2016. “Ritual Human Sacrifice Promoted and Sustained the Evolution of Stratified Societies.” Nature 532 (7598): 228–31. https://doi.org/10.1038/nature17159.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” Scientific Data 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.