3 From Old Data to Fresh Phylogenies — A Linguistic Data Journey in the Times of CLDF

Christoph Rzymski[1]

[1] Dep. of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology Leipzig

Historical linguistics involves the study of language change over time, and is often aided by the use of cross-linguistic data. Cross-Linguistic Data Formats (CLDF, Forkel et al. 2018) provides a standardized way to represent and exchange such data, while cldfbench (Forkel & List 2020) is a workflow model that facilitates the management and analysis of CLDF data. In this study, we demonstrate how CLDF and cldfbench can be used to tackle commonplace tasks in historical linguistics, such as analyzing word lists to identify cognates and building phylogenies. By using CLDF as both input and output, we aim to show how these tools can help streamline the process of working with cross-linguistic data in historical linguistics, from the initial stage of collecting data from “old sources” (i.e., physical sources such as dictionaries and language documentation materials) to the final stage of constructing phylogenies that represent the relationships between languages.

We will demonstrate how to automatically compute cognates (List 2018, List 2021) in word lists using resources such as Concepticon (List 2022) and Glottolog (Hammarström 2022), and how to use these lists as input for BEAST (Bouckaert et al. 2014) to compute phylogenies. Since cldfbench supports a workflow that involves using “raw” source data and converting it to one or more CLDF datasets with the help of custom configurations and/or additional Python code, we aim to showcase how this can be utilized to prepare datasets for individual research questions. CLDF, cldfbench, and the aforementioned workflows can help researchers to efficiently process and analyze large amounts of data, and facilitate the integration of data from multiple sources.

Overall, our goal is to demonstrate the utility of CLDF and CldfBench for researchers in the field of historical linguistics, and to encourage their adoption as standard tools for handling cross-linguistic data. By showcasing innovative approaches to working with standardized cross-linguistic data, we hope to inspire new ideas and perspectives on how to build fresh phyologenies from “old data”.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.H., Xie, D., Suchard, M., Rambaut, A., & Drummond, A. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Computational Biology, 10(4), e1003537.

Forkel, R., List, J.M., Greenhill, S., Rzymski, C., Bank, S., Cysouw, M., Hammarstrom, H., Haspelmath, M., Kaiping, G., & Gray, R. (2018). Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. Scientific data, 5(1), 1–10.

Forkel, R., & List, J.M. (2020). CLDFBench. Give your Cross-Linguistic data a lift. In Proceedings of the Twelfth International Conference on Language Resources and Evaluation (pp. 6997-7004). European Language Resources Association (ELRA).

Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2022). CLLD Glottolog 4.7. Max Planck Institute for Evolutionary Anthropology. https://glottolog.org.

List, J. M., Walworth, M., Greenhill, S. J., Tresoldi, T., & Forkel, R. (2018). Sequence comparison in computational historical linguistics. Journal of Language Evolution, 3(2), 130-144.

List, J.M, & Forkel, R. (2021). LingPy. A Python library for historical linguistics. Version 2.6.9. URL: https://lingpy.org, DOI: https://zenodo.org/badge/latestdoi/5137/lingpy/lingpy. With contributions by Greenhill, Simon, Tresoldi, Tiago, Christoph Rzymski, Gereon Kaiping, Steven Moran, Peter Bouda, Johannes Dellert, Taraka Rama, Frank Nagel. Leizpig: Max Planck Institute for Evolutionary Anthropology.

List, J.M., Tjuka, A., Rzymski, C., Greenhill, S., & Forkel, R. (2022). CLLD Concepticon 3.0.0. Max Planck Institute for Evolutionary Anthropology. https://concepticon.clld.org.