Open data in Computational Historical Linguistics: Upcycling of Chibchan wordlists

Frederic Blum
Humboldt-Universität zu Berlin
frederic.blum@hu-berlin.de

Keywords: Computational Historical Linguistics, Open Science, Phylolinguistics, Chibchan.

Abstract

This work presents presents an ongoing research project on the phylogeny of the Chibchan language family which adheres to the current state-of-the-art of computational methods in Historical Linguistics. The aim is to show how we can upcycle already published data and infer new results by applying models of Bayesian inference (or others). The following steps of the workflow will be introduced:

i. Collecting already published data for the Chibchan family (Constenla Umaña 2005)
ii. Applying cross-linguistic data format standards (Forkel & List 2020)
iii. Detection of cognacy and working with Edictor (List 2017)
iv. phylogenetic analysis through Bayesian inference (Greenhill et al. 2020)

All data of this project was published by Constenla Umaña (2005) and will be upcycled into the Cross-Linguistic Data Format (Forkel & List 2020) which adheres to certain standards which ensure comparability across languages. In the next step, this data is first analyzed automatically for cognates and then manually corrected where necessary (List 2017). In a final step, the cognates will be analyzed with computational methods, such as NeighborNetworks and Bayesian inference of phylogenetic trees (Greenhill et al. 2020).

The main goal of this presentation is to show how computational methods can enrich our workflows also in more 'traditional' areas of linguistics and how we can upcycle old data to arrive at new conclusions about the languages of the world. In times of a pandemic, these seem to be reasonable methods to gather data when we cannot do fieldwork and the data can be used for a huge variety of goals. All data and scripts used in this presentation will be published in GitHub and/or Zenodo.

References

Constenla Umaña, Adolfo. 2005. ¿existe relación genealógica entre las lenguas misumalpas y las chibchenses? *Estudios de Lingüística Chibcha* .

Forkel, Robert & Johann-Mattis List. 2020. Cldfbench. give your cross-linguistic data a lift. In *Proceedings of the twelfth international conference on language resources and evaluation.* 6997-7004. Luxembourg: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.864.pdf.

Greenhill, Simon J, Paul Heggarty & Russell D Gray. 2020. Bayesian phylolinguistics. *The Handbook of Historical Linguistics* 2. 226–253.

List, Johann-Mattis. 2017. Historical language comparison with lingpy and edictor. *Jena: Max Planck Institute for the Science of Human History, Linguistic and Cultural Evolution. https://github.com/digling/edictor-tutorial* .