Automated Phonological Reconstruction Based on Sound Correspondence Patterns

Nathan W. Hill[1] and Johann-Mattis List[2]

[1] Dublin University [2] Max Planck Institute for Evolutionary Anthropology

Computational approaches in historical linguistics have been increasingly applied during the past decade and many new methods that implement parts of the traditional comparative method have been proposed. Despite these increased efforts, there are not many easy-to-use and fast approaches for the task of automated phonological reconstruction, in which an algorithm would have to infer the best reconstructions for a larger number of cognate sets.

Recent automatic approaches for linguistic reconstruction, be they supervised (working with training data, in which valid reconstructions are given as examples for the algorithm to learn) or unsupervised (working without any form of training data and giving the algorithm only the raw data), show two major problems. First, the underlying code is rarely made publicly available, which means that they cannot be further tested by applying them to new datasets. Second, the methods have so far only been teste on a small amount of data from a limited nuber of language families. Thus, Bouchard-Côté et al. (2013) report remarkable results on the reconstruction of Oceanic languages, but the source code has never been published, and the method was never tested on additional datasets. Meloni et al. (2021) report very promising results for the automated reconstruction of Latin from Romance languages, using a new test set derived from a dataset originally provided by Dinu and Ciobanu (2014), but their source code has not been shared in a state that it could be easily applied to other datasets and only part of the data could be shared, since the authors from which they received the data did not allow them to share them. Bodt and List (2021) experiment with the prediction of so far unelicited words in a small group of Sino-Tibetan languages, but they do not test the suitability of their approach for the reconstruction of ancestral languages. Jäger (2019) presents a complete pipeline by which words are clustered into cognate sets and ancestral word forms are reconstructed, but the method is only tested on a very small dataset of Romance languages.

In our talk, we present a new framework for automated linguistic reconstruction which combines state-of-the-art methods for automated sequence comparison with fast machine-learning techniques and test it on a newly compiled test set that covers multiple language families.

The new framework can be divided into a training and a prediction stage. The training consists of four steps. In step (1), the cognate sets in the training data are aligned with a multiple phonetic alignment algorithm. In step (2), the alignments are trimmed by merging sounds in the ancestral language into clusters which would leave no trace in the descendant languages. In step (3), the alignments of the descendant languages are enriched by coding for context that might condition sound changes. In step (4) the enriched alignment sites are assembled and fed to a classifier for training.

Our tests show that our framework yields quite reliable, albeit not perfect, reconstructions for a diverse set of languages using data that comes along in different sizes. Given that data and code are publicly available and can be easily applied without having too much programming experience, our approach can already be useful for scholars who work on so far understudied language families in helping them to scale up the annotation of cognate sets.

References

Timotheus Adrianus Bodt and Johann-Mattis List. 2021. Reflex prediction. a case study of Western Kho-Bwa. Diachronica, 0(0):1–38.

Alexandre Bouchard-Côté, David Hall, Thomas L. Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. Proceedings of the National Academy of Sciences of the United States of America, 110(11):4224–4229.

Liviu Dinu and Alina Maria Ciobanu. 2014. Building a dataset of multilingual cognates for the Romanian lexicon. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 1038–1043, Reykjavik, Iceland. European Language Resources Association (ELRA).

Carlo Meloni, Shauli Ravfogel, and Yoav Goldberg. 2021. Ab antiquo: Neural proto-language reconstruction. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4460–4473, Online. Association for Computational Linguistics.