Using simulated data to evaluate models of Indo-European vocabulary evolution

Philipp Rönchen[1], Oscar Billing[1], and Tilo Wiklund[2]
1Department of Linguistics and Philology, Uppsala University, Sweden
2Chief Data Scientist, UAB Sensmetry, Vilnius, Lithuania

In the last two decades the project of using data from the lexicon of modern languages to make inferences about historical language stages, though long envisioned (Hymes 1960, Embleton 1986), has been gaining steam. Gray and Atkinson (2003), Bouckaert et al. (2012) and Chang et al. (2015) use increasingly sophisticated methods to estimate the age of Indo-European, however the results of the earlier studies run counter to the established majority opinion in historical linguistics (Pronk, 2022) and Chang et al.'s methodology gives a different result. This raises the question how different computational models can be validated (see Nakhleh et al. 2005, Ritchie and Ho 2019, Jäger 2019a and 2019b)

Ideally one would like to evaluate computational methods using held-out data sets and test cases in which the correct inferences are known. However, compared to other disciplines like biology, the amount of lexical data available in data bases is very limited and the precise history of most language families in the world is unknown, leaving only a few quite shallow families as potential test cases. Moreover, it is not clear whether the success of a computational model on a language family from one part of the world should generalise to other families, since different evolutionary mechanisms might have operated. To work around the lack of data available for validation, Greenhill et al. (2009), Murawaki (2015) and Bradley (2016) simulate data sets which they use to evaluate computational methods.

We create a large number of simulated data sets to evaluate the inferences of Chang et al. (2015) and Bouckaert et al. (2012) on Indo-European. Our data sets are specifically tailored to the methodologies of Chang et al. and Bouckaert et al. and try to mimic different plausible (though hypothetical) pre-histories of Indo-European, including loan events, a tree topology not too far from the consensus view in historical linguistics, and varying lexical change rates. We employ the computational fact that it is much easier to create realistic models for simulating data then it is to make inferences from existing data (see Kelly and Nicholls 2017 for the difficulties involved in constructing an inference method that allows for loans).

Both Chang et al.'s and Bouckaert et al.'s methodologies fail to correctly infer the age of Indo-European that was used to create our simulated data sets. We believe this warrants more investigation in the validity of different computational models.

References

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A., and Atkinson, Q. D. (2012). Mapping the origins and expansion of the Indo-European language family. Science, 337(6097):957–960.

Bradley, S. (2016). Synthetic language generation and model validation in BEAST2. arXiv preprint arXiv:1607.07931.

Chang, W., Cathcart, C., Hall, D., and Garrett, A. (2015). Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. Language, 91(1):194–244.

Embleton, S. M. (1986). Statistics in historical linguistics, volume 30. Brockmeyer.

Gray, R. D. and Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature, 426(6965):435–439.

Greenhill, S. J., Currie, T. E., and Gray, R. D. (2009). Does horizontal transmission invalidate cultural phylogenies? Proceedings of the Royal Society B: Biological Sciences, 276(1665):2299–2306.

Hymes, D. H. (1960). Lexicostatistics so far. Current anthropology, 1(1):3–44.

Jäger, G. (2019a). Computational historical linguistics. Theoretical Linguistics, 45(3-4):151–182.

Jäger, G. (2019b). Model evaluation in computational historical linguistics. Theoretical Linguistics, 45(3-4):299–307.

Kelly, L. J. and Nicholls, G. K. (2017). Lateral transfer in stochastic dollo models. The Annals of Applied Statistics, 11(2):1146–1168.

Murawaki, Y. (2015). Spatial structure of evolutionary models of dialects in contact. Plos one, 10(7):e0134335.

Nakhleh, L., Warnow, T., Ringe, D., and Evans, S. N. (2005). A comparison of phylogenetic reconstruction methods on an IE dataset. Transactions of the Philological Society, 3(2):171–192.

Pronk, T. (2022). Indo-European secondary products terminology and the dating of Proto-Indo-Anatolian. Journal of Indo-European Studies, 49(1&2):141–170.

Ritchie, A. M. and Ho, S. Y. (2019). Influence of the tree prior and sampling scale on Bayesian phylogenetic estimates of the origin times of language families. Journal of Language Evolution, 4(2):108–123.