Tongues, trees and Bayesian Inference: towards a global phylogeny of the world's languages

Quentin D. Atkinson[1,2]
[1] School of Psychology, University of Auckland
[2] Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology

Keywords: Computational modelling, Cultural evolution, Linguistic diversity

Over the last two centuries, the field of historical linguists has enabled increasingly detailed inferences about ancestral relationships between the world's languages, identifying more than 200 language families and almost as many isolates (1). However, most linguists consider a global 'tree of language', akin to biology's 'tree of life', out of reach, and view proposed language 'super-families', such as Amerind (2), Nostratic (3, 4) and Eurasiatic (5, 6), with extreme skepticism. There are good reasons for this. First, high rates of language change tend to 'wash out' deep historical signal in the data and make it difficult to reliably distinguish true homologies from chance resemblances or more recent borrowing (3, 7–9). Second, without a probabilistic model of how languages change through space and time, there is no statistical framework with which to compare multiple lines of evidence or evaluate competing hypotheses. And third, the universe of possible global trees is so vast that manually comparing the alternatives becomes intractable - for just 100 languages, there are more possible trees than there are atoms in the universe (10).

Nevertheless, in recent years, newly available data and methods have given rise to renewed interest in deep connections between languages and the possibility of pushing back the time depth of historical linguistic inference (7, 11–24). Innovations include the development of new cross-linguistic databases (15, 19, 23), explicitly targeting highly conserved traits (11, 12), new tools for phylogenetic inference and modelling language change (12, 18, 24), and integrating genetic and linguistic data (15, 20–22).

In this talk I will review the state of the art and argue that while each of the existing approaches has its weaknesses, together this body of work has the potential to overcome many of the limitations of earlier studies. I will then showcase new data emerging from the Glottobank project (www.glottobank.org) and a novel Bayesian phylogeographic approach implemented in BEAST (25), designed to integrate findings from diverse data and methodologies into a principled statistical framework. I will show how the resulting posterior distribution of trees can be used to clarify both what we can and, crucially, cannot say about the origins of human linguistic diversity, an application that should appeal to both proponents and critics of 'long range' language relationships.

Finally, I will show how, despite considerable phylogenetic uncertainty inherent in existing knowledge, the posterior distribution of trees representing this knowledge can nevertheless be used to identify key social and ecological drivers of linguistic diversity around the globe. I conclude by discussing future applications of this work to questions concerning human prehistory, the evolution and conservation of linguistic diversity, and the evolution of human culture more broadly.

References

1. H. Hammarström, R. Forkel, M. Haspelmath, Glottolog 4.0 (2019).
2. J. H. Greenberg, *Language in the Americas* (Stanford University Press, 1987).
3. R. L. Oswalt, A probabilistic evaluation of North Eurasiatic Nostratic. Amsterdam Studies in the Theory and History of Linguistic Science, 4, 199–216 (1998).
4. A. Dolgopolsky, *The Nostratic Macrofamily and Linguistic Palaeontology* (McDonald Institute for Archaeological Research, 1998).
5. M. Ruhlen, *A Guide to the World's Languages: Classification* (Stanford Univ, 1987).
6. J. H. Greenberg, *Indo-European and Its Closest Relatives: The Eurasiatic Language Family, Volume 2, Lexicon* (Stanford University Press, 2000).
7. R. Gray, Evolution. Pushing the time barrier in the quest for language roots. *Science* **309**, 2007–2008 (2005).
8. D. A. Ringe Jr, "Nostratic" and the factor of chance. *Diachronica* **12**, 55–74 (1995).
9. L. Campbell, W. J. Poser, *Language classification: History and method* (Cambridge University Press, 2008).
10. J. Felsenstein, J. Felenstein, *Inferring phylogenies* (Sinauer Ass. Sunderland, MA, 2004).
11. M. Dunn, A. Terrill, G. Reesink, R. A. Foley, S. C. Levinson, Structural phylogenetics and the reconstruction of ancient language history. *Science* **309**, 2072–2075 (2005).
12. M. Pagel, Q. D. Atkinson, A. S Calude, A. Meade, Ultraconserved words point to deep language ancestry across Eurasia. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 8471–8476 (2013).
13. M. A. Sicoli, G. Holton, Linguistic phylogenies support back-migration from Beringia to Asia. *PLoS One* **9**, e91722 (2014).
14. I. Yanovich, Phylogenetic linguistic evidence and the Dene-Yeniseian homeland. *Diachronica* **37**, 410–446 (2020).
15. M. Robbeets, *et al.*, Triangulation supports agricultural spread of the Transeurasian languages. *Nature* **599**, 616–621 (2021).
16. G. Jäger, Global-scale phylogenetic linguistic inference from lexical resources. *Sci Data* **5**, 180189 (2018).
17. D. Dediu, Making genealogical language classifications available for phylogenetic analysis: Newick trees, unified identifiers, and branch length. *Language Dynamics and Change* **8**, 1–21 (2018).
18. Q. D. Atkinson, Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* **332**, 346–349 (2011).
19. G. Longobardi, C. Guardiano, Evidence for syntax as a signal of historical relatedness. *Lingua* **119**, 1679–1706 (2009).
20. G. Longobardi, *et al.*, Across language families: Genome diversity mirrors linguistic variation within Europe. *Am. J. Phys. Anthropol.* **157**, 630–640 (2015).
21. P. Duda, J. Zrzav\`y, Towards a global phylogeny of human populations based on genetic and linguistic data. *Modern Human Origins and Dispersal. Words, Bones, Genes, Tools: DFG Center for Advanced Studies Series* **1**, 331–359 (2019).
22. P. Duda, Jan Zrzavý, Human population history revealed by a supertree approach. *Sci. Rep.* **6**, 29890 (2016).
23. J. Dellert, *et al.*, NorthEuraLex: a wide-coverage lexical database of Northern Eurasia. *Language Resources and Evaluation* **54**, 273–301 (2020).
24. G. Jäger, Support for linguistic macrofamilies from weighted sequence alignment. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 12752–12757 (2015).
25. R. Bouckaert, *et al.*, BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **15**, e1006650 (2019).