A New Database, Family Tree and Origins Hypothesis for the Indo-European Language Family

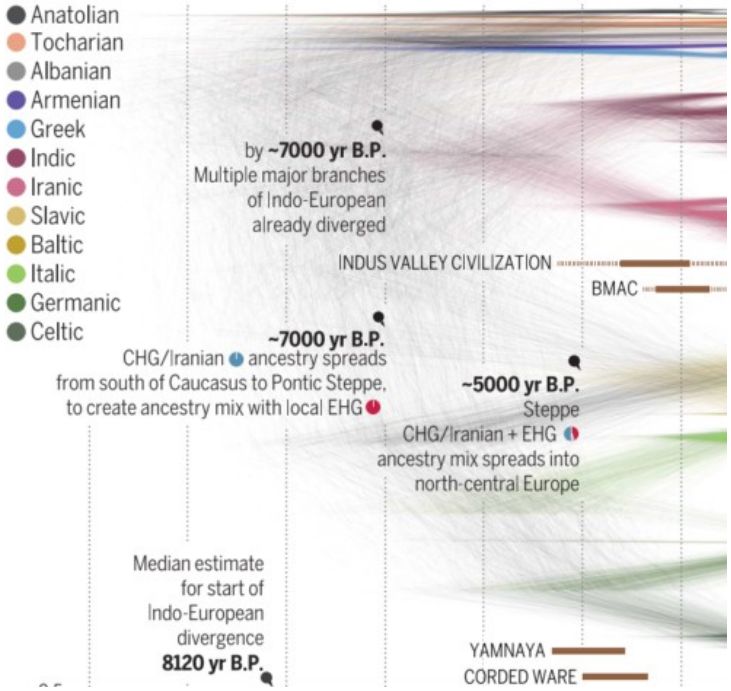
Paul Heggarty

Language and the Anthropocene Group, Max Planck Institute for Geoanthropology, Jena, Germany

Dpto de Humanidades, Pontificia Universidad Católica del Perú

This talk reports on a new Bayesian 'phylo-chronology' of humanity's most widely spoken language lineage. The results have reignited controversy over the age and origins of the Indo-European family — and over whether we can really trust such phylogenetic analysis methods to reliably recover how 'evolution' proceeds specifically in *language*.

I present first the new IE-CoR language database — useful independently for many other research ends, too. I show how the phylogenetic analysis came to its results, in tree structure and chronology, and how those point to a new 'hybrid' hypothesis of Indo-European origins, assessed also against archaeology, ancient DNA, and some widely-held objections. I clarify some ongoing misconceptions about Bayesian phylogenetics, and how to read its results — but also illustrate outstanding issues that the methodology still faces, and possible next steps to attenuate them.



More … ↓

The *Science* paper on the new 'IE-CoR' phylogenetic analysis of Indo-European is available from the free link at iecor.clld.org. Further resources and background on controversial issues can be found at the support page at paulheggarty.info/. In presenting this research, this talk will cover five main topics.

- The Indo-European Cognate Relationships database of 161 languages (iecor.clld.org), and its multiple new departures in database methodology, devised to serve various research aims in quantitative comparative linguistics.
- The results of Bayesian phylogenetic analyses of the IE-CoR data: the tree structure, and estimated chronology that Indo-European began to spread and diverge from c. 8100 years ago (95% highest posterior density: 9610 to 6740 BP).
- What this chronology, tree topology, and the place and age of Indo-Iranic, suggest for Indo-European origins: neither the Steppe nor farming hypothesis, but a 'hybrid' of both, as assessed also against archaeology and ancient DNA.
- Ongoing misconceptions on this methodology: what splits represent; how the concepts of 'a language', registers and dialects map onto them; and how cognacy trees can differ from qualitative phonology and/or morphology trees.
- Basic outstanding issues facing Bayesian phylogenetic methodology, and possible next steps to attenuate them.

Paul Heggarty is a historical linguist with a focus on how our languages open up a unique window on our past. He works and publishes explicitly in concert with other disciplines, especially archaeology and human genetics, for a more coherent, holistic understanding of the human past, across all those perspectives. His career has spanned research institutes in linguistics, archaeology, human history, evolutionary anthropology and geoanthropology, at Cambridge and in the Max Planck system. He founded and runs www.soundcomparisons.com and led the IE-CoR database on cognate relationships across Indo-European (iecor.clld.org), revealing both the strengths and weaknesses of Bayesian phylogenetics as applied to language families. His interests range worldwide, but focus also on South America (see paulheggarty.info/symposia), where he currently holds a *cátedra de excelencia* at the Pontificia Universidad Católica del Perú in Lima.