Testing hypotheses of Evolutionary Biology in linguistics

Κωνσταντίνος Γεώργιος Σταμέλος
Επιβλέπων: Εμμανουήλ Λαδουκάκης

Background

- Charles Dawrin in his book “The descent of man” (1872) states that “The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously the same”.
- Approaching languages as an evolvable system allows for the testing of hypotheses that come from evolutionary biology. One of the most principal among them is whether characters and species evolve at a constant rate (Ladoukakis et al., 2022).
- **Goal:** to test whether the rates of evolution across languages are constant or not, focusing on the Indo-European family tree (IE).

Approach

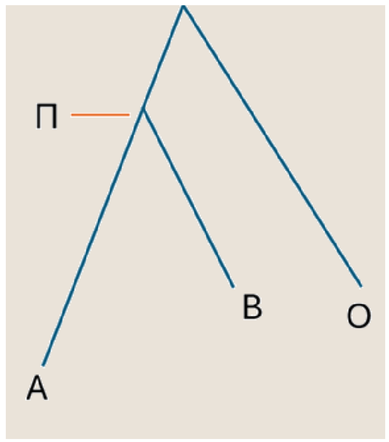
The analysis was based on a curated dataset of morphosyntactic features (Sleeman et al., 2026), with languages as rows and features as columns. The data consisted of 113 features for 61 languages of the Indo-European family

➢ Method of distances

To quantify similarity between languages, the pairwise linguistic distances were computed using Gower’s distance in a triadic framework using using R version 4.5.1 and Rstudio version 2025.091 build 401. Each time, two focal languages (A, B) were compared to a third language acting as an outgroup (O). For each unique triad (A, B, O) the difference “D” between the distance from A to O and from B to O was estimated, as shown in figure 1.

For each language pair (A, B), multiple outgroups (O) were used to evaluate relative divergence based on distance differences. Evidence was aggregated across every meaningful outgroup of each pair of interest (based on a known IE phylogeny). A normalized duo-level dominance index was defined as follows:

$$D_{AB} = \frac{H_{A>B} - H_{B>A}}{N_{AB}}$$



A	B	N_{AB}	$H_{A>B}$	$H_{B>A}$
Language 1	Language 2	8	7	1
Language 2	Language 3	9	1	8
Language 1	Language 3	10	5	5

To further summarize relative rate patterns at the level of individual languages, duo-level dominance indices were combined across all relevant comparisons into a **language specific index Index(L’)**. Index(L’) was estimated using random-effects meta-analysis; the DerSimonian-Laird estimator (DerSimonian & Laird, 2015), accounting for sample heterogeneity between pairwise comparisons and for outgroups that belonged to closely-related languages and were therefore considered pseudo-replicates.

➢ Bayesian Approach

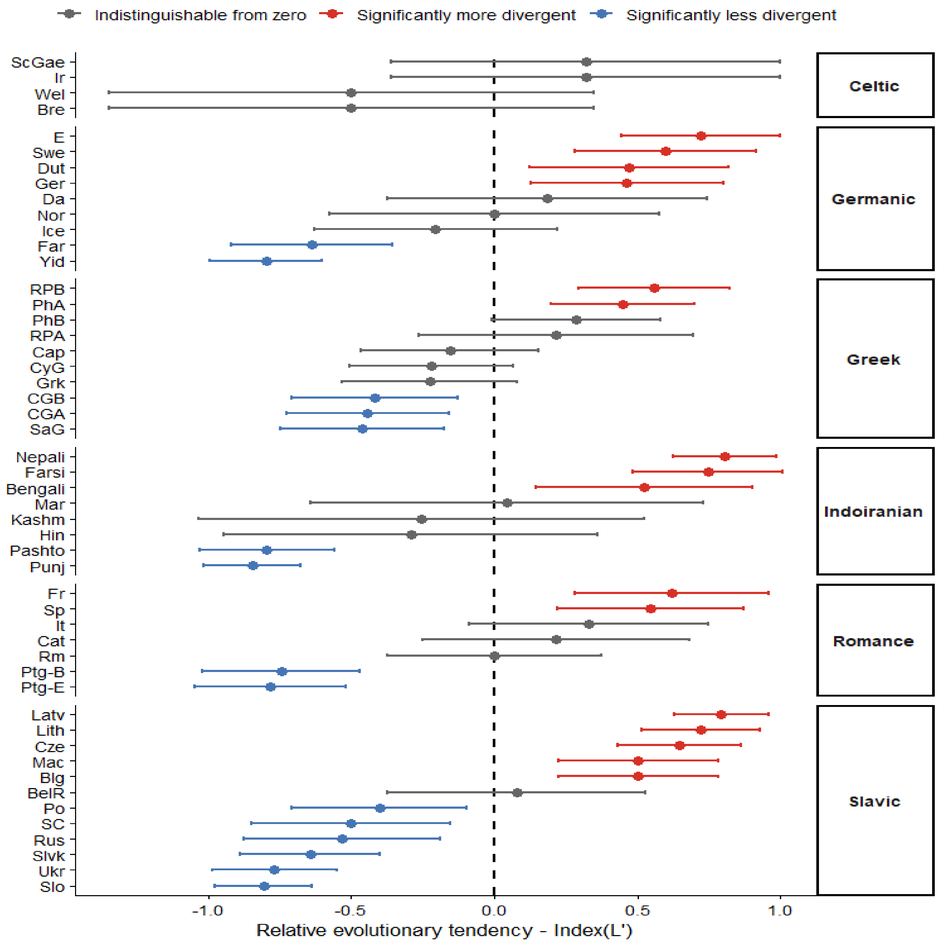
BEAST 2 version 2.7 (Bouckaert et al., 2019) and package BEASTlabs was used to quantify the evolutionary rates in the whole phylogenetic tree of the IE languages that were part of this analysis. Characters were modeled under a Lewis-MK model of discrete rate evolution and data were partitioned according to the number of observed states for every character. A known phylogeny was implemented using BEAST’s “fixed tree analysis” (Bouckaert, 2023) in order to account for the time parameter in the branch length = rate x time equation. This way, branch specific rates could be inferred. The Uncorrelated LogNormal Relaxed Clock was used to do so, and also the Strict Clock for reference.

➢ Simulated Data

To test if these data could be fruitful in a Bayesian phylogenetic analysis and could support this kind of inference, simulated data were created. Using a python script pipeline developed by colleagues Pavlos Pavlidis & Christos Zioutis, phylogenies were simulated from Heggarty’s tree of the Indo-European languages (Heggarty et al., 2023). Simulated data sequences produced were used to generate XML files that would later be run in BEAST. Evolutionary rates of this data were known from the simulation. Therefore, it could be tested by the results if they are informative. In a similar way, they were tested with the distance approach for verification of this method.

Results

For the distance methods, the average directional distance asymmetry of each language within its clade is shown below.



Discussion

The evolutionary rates that index(L’) represents convey the message that a given language has diverged more, or not, compared to its close relatives. Owing to the contemporaneity of all languages studied, divergence translates to evolutionary rate. **It is supported that morphosyntactic change is accumulating in lineage-specific ways that are detectable from analytical perspectives and that rate heterogeneity is an intrinsic evolutionary characteristic of languages in the level of morphosyntax.** Technical problems concerning the Bayesian approach made it difficult to infer evolutionary rates and will be dealt with in the future, to cross-examine the results of both methods.

References:

Bouckaert, R. (2023). *FixedTreeAnalysis: A BEAST 2 package for performing an analysis where the tree is fixed* (Version X.X.X). GitHub. https://github.com/rbouckaert/FixedTreeAnalysis
DerSimonian, R., & Laird, N. (2015). Meta-analysis in clinical trials revisited. *Contemporary Clinical Trials*, *45*, 139–145. https://doi.org/10.1016/j.cct.2015.09.002
Heggarty, P., Anderson, C., Scarborough, M., King, B., Bouckaert, R., Jocz, L., Kümmel, M. J., Jügel, T., Irslinger, B., Pooth, R., Liljegren, H., Strand, R. F., Haig, G., Macák, M., Kim, R. I., Anonby, E., Pronk, T., Belyaev, O., Dewey-Findell, T. K., … Gray, R. D. (2023). Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages. *Science*, *381*(6656). https://doi.org/10.1126/science.abg0818
Ladoukakis, E. D., Michelioudakis, D., & Anagnostopoulou, E. (2022). Toward an evolutionary framework for language variation and change. *BioEssays*, *44*(3). https://doi.org/10.1002/bies.202100216
Sleeman, R., Makri, M.-M., Anagnostopoulou, E., Ladoukakis, E. D., Michelioudakis, D., Zioutis, C., & Pavlidis, P. (2026). Evaluating the phylogenetic signal of morphosyntax. *Poznan Studies in Contemporary Linguistics*. https://doi.org/10.1515/psicl-2025-0030