How time-bound is your grammar?
Pushing the limits of comparative methods through syntax

L. Bortolussi, A. Ceolin, G. Cordoni, C. Guardiano, D. Kazakov,
G. Longobardi, M. A. Irimia, N. Radkevich, A. Sgarro

Introduction A long-standing probabilistic question in historical linguistics is how much similarity is required to demonstrate true language relatedness in the absence of regular sound correspondences. Attempts to find significant phonetic similarities between languages across the traditionally established language families in terms of vocabulary items have so far failed to uncontroversially solve this problem. Crucially, many methods retrieve phylogenies on quantitative grounds (Ringe et al 2002, Gray and Atkinson 2003) and some even assume macro-families without testing whether the similarities they consider are safe against a null hypothesis of unrelatedness (Pagel et al. 2013, Jäger 2015). Here we shift the focus from the vocabulary to syntax by using universally definable parameters of generative grammar as taxonomic characters. From such characters, we calculate language distances and phylogenetic trees (**Figure 1**). Then, we develop an algorithm that generates theoretically possible grammars to perform a probabilistic testing of taxonomic hypotheses. We calculate pairwise syntactic distances within a random sample of such simulated languages and plot their distribution against that of syntactic distances calculated within several sets of actual languages falling into different attested and proposed language families (**Figure 2**).

Results Using this mathematical model, we first show that the distributions of syntactic distances within each of the Indo-European (Median, M=0.259), Uralic (M=0.273), and Altaic families (M=0.2) are statistically significantly different ($p<0.00001$, according to a Mood's scale test) from the ones simulated by the random generation (M=0.545, Figure 2). Then, we apply the same procedure to the controversial Ural-Altaic macro-family and obtain equally positive evidence ($p<0.00001$). Importantly, other groupings of languages that have been proposed in the literature, such as Indo-Uralic, instead fail to be supported by this method ($p=0.2571$). We also apply several correlation tests between our syntactic distances and geographical distances (GCD) to control for the effect of areal convergence. Such tests show that our cross-family distance pairs are not correlated with geographical ones, therefore ruling out horizontal transmission as the main source of linguistic similarity.

Conclusion Syntax has been regarded in both the historical and the recent generative tradition as unable to provide insights about language relatedness. Contrary to this widespread bias, we demonstrate that, in fact, syntax contains a detectable historical signal and that, unlike classical methods, it can provide time-deep phylogenies encompassing distinct traditional families. Our results show that the deep and controversial Ural-Altaic cluster is statistically supported, while other possible macro-groupings like e.g. Indo-Uralic, sometimes proposed in the literature, fail the same test. Finally, with the help of Bayesian character methods, we will attempt to hypothesize a first reconstruction of the syntactic states of the Ural-Altaic protolanguage.

References Gray, R. D., & Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, *426*(6965), 435-439. Jäger, G. (2015). Support for linguistic macrofamilies from weighted sequence alignment. *Proceedings of the National Academy of Sciences*, *112*(41), 12752-12757. Pagel, M., Atkinson, Q. D., Calude, A. S., & Meade, A. (2013). Ultraconserved words point

to deep language ancestry across Eurasia. *Proceedings of the National Academy of Sciences*,*110*(21), 8471-8476. Ringe, D., Warnow, T. & Taylor, A. Indoeuropean and computational cladistics. *Transactions of the Philological Society 100*(1), 59-129.

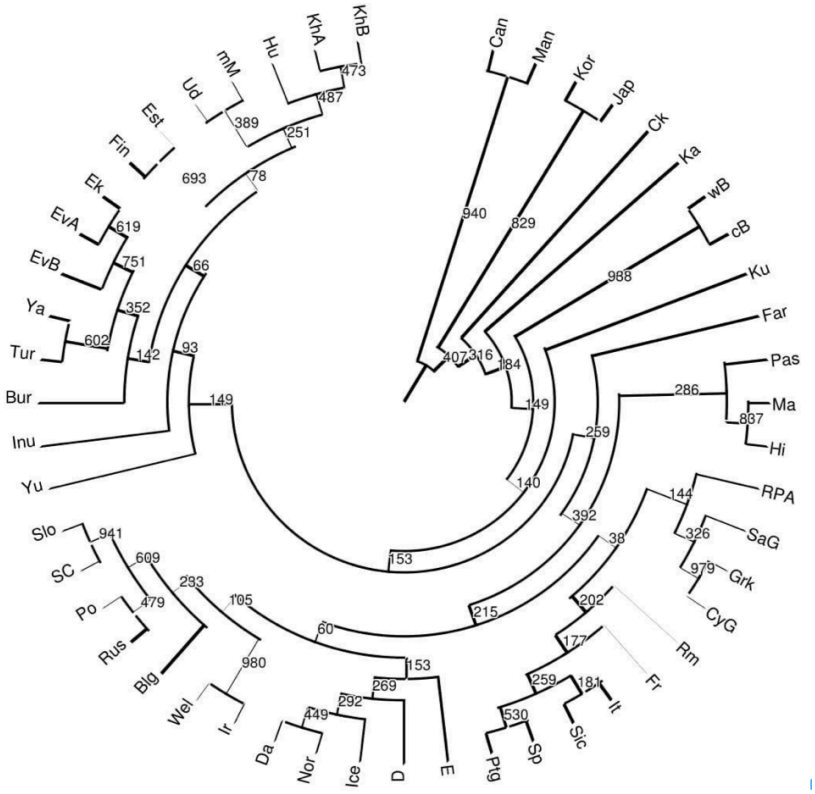

Figure 1: Phylogenetic tree generated by Kitsch, after a bootstrapping procedure

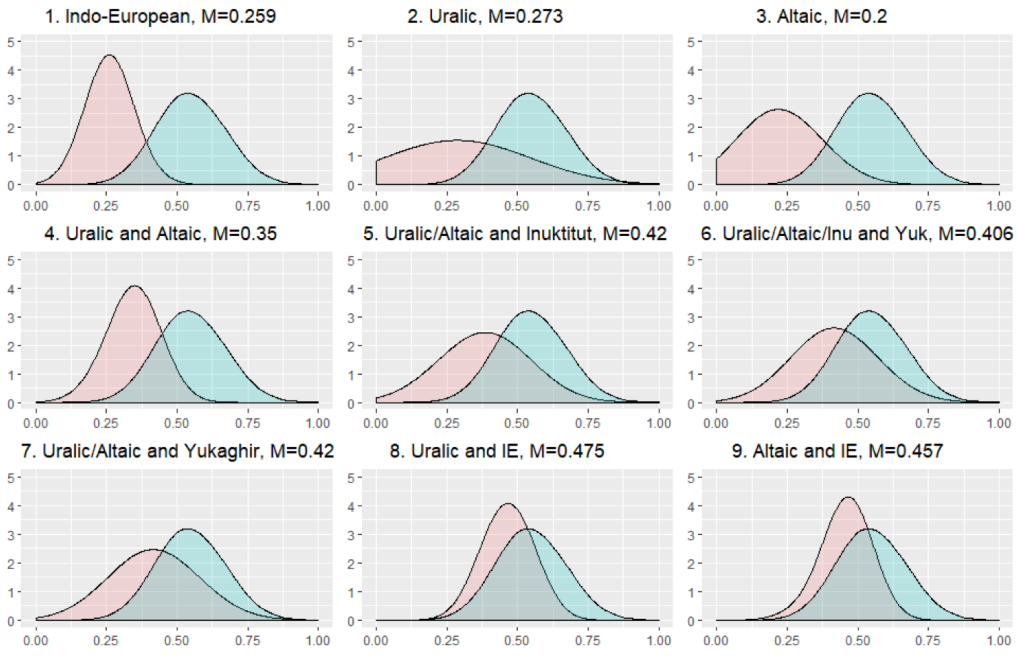

Figure 2: Histograms of various proposed families plotted against our simulated sample