Generative syntax and the linguistic prehistory of Eurasia

For over 200 years syntax has been excluded from the tools of language taxonomists under the presupposition that syntactic diversity does not contain a reliable signal of historical language relatedness, a position explicitly maintained even after the development of generative diachronic studies (Anderson&Lightfoot 2002:8-9). In this presentation, we provide quantitative and statistical evidence that demonstrates, against this skepticism, that syntax retains a historical signal able to suggest time-deep language relations. We use the Parametric Comparison Method (PCM, Longobardi&Guardiano 2009, Ceolin et al. 2020) to explore cross-family relations in Eurasia, by combining phylogenetic reconstruction methods with a theory of possible grammars.

We employ a dataset containing values for 94 syntactic nominal parameters in 58 languages belonging to 15 established families (Fig.1). First, we calculate their pairwise distances by means of a Jaccard metric and we analyze their distribution through standard quantitative procedures (Heatmaps, PCoAs, phylogenetic algorithms): the language aggregations obtained from syntactic distances are almost completely consistent (nearly 90%) with the vertical etymological signal independently defining families and subfamilies.

To test the statistical robustness of the nodes/aggregations which do not correspond to traditional historical knowledge (or correspond to more controversial classifications), we apply a dedicated statistical test inspired by the literature on significant testing of language relatedness (Oswalt 1970, Ringe 1992, Kessler&Lehtonen 2006). Elaborating on Bortolussi et al. (2011), we generate the class of theoretically possible languages predicted by our parameter system, taking into account its implicational structure; then, we compare the Jaccard distances derived from a sample of such languages with those of our dataset, to determine whether they have a similar distribution. The test identifies a significance threshold of d=0.33, under which distances are unlikely to appear by chance only.

Jaccard distances drawn from languages belonging to the same established family typically fall below the significance threshold. By contrast, the distances between languages which do not belong to the same established family are generally higher than 0.33, and therefore uninformative, with four remarkable exceptions: (1) the distance between Korean and Japanese (d=0.182, p=0.003), (2) the mean distance between NE Caucasian and Dravidian (d=0.263, p=0.024), (3) the mean distances between the languages belonging to the so-called (micro)Altaic group (Buryat, Tungusic, and Turkic), and (4) the mean distances between the latter and Finno-Ugric. Thus, these results call for a historical explanation.

We find some independent support for (1) and (2) in the evidence provided by population genetics and ancient DNA studies (Reich 2019). As for (1), the syntactic proximity of Japanese and Korean matches the similarity in the genetic profiles of the two corresponding modern populations more than their still controversial lexical-etymological relation. As for (2), there are proposals, based on linguistic evidence, suggesting the existence of a Proto-Elamo-Dravidian family, which connects Dravidian with Elamite, a now extinct language spoken in a territory roughly corresponding to western Iran, which in turn shares geographic borders and some non-trivial linguistic properties with Caucasian languages: in other words, Elamite seems to be a potential link that indirectly connects Dravidian and NE Caucasian, a suggestion that matches our findings. Even more interestingly, genetic studies also show some relatedness between modern Dravidian speakers of South India and of ancient Iranian farmers (11000-8000 BC; Lazaridis et al. 2016, Reich 2019), who occupied an area that is in turn geographically close, and in exchange relationships, with the Caucasus. As for (3) and (4), comparison between genes and syntactic parameters (Santos et al. 2020) has shown that Finno-Ugric and Altaic speakers cannot be traced back to a common genetic

pool for independent demographic and historical reasons, so that linguistic insights are the best contribution to the understanding of their prehistory.

While our test suggests a positive result for micro-Altaic and even for a larger Uralo-Altaic, it provides no evidence of other controversial superfamilies discussed in the macro-comparative literature (e.g. Indo-Uralic, Sino-Caucasian, Basque-Caucasian, macro-Altaic).

These results suggest that the generative modelling of syntactic diversity can be used to: (a) provide a proof of historical relation between different families irrespectively of the presence regular sound correspondences, thus expanding the time limits imposed by traditional comparative methods; (b) open new perspectives for applying the discoveries and methods of biolinguistics and cognitive science to historical anthropology.

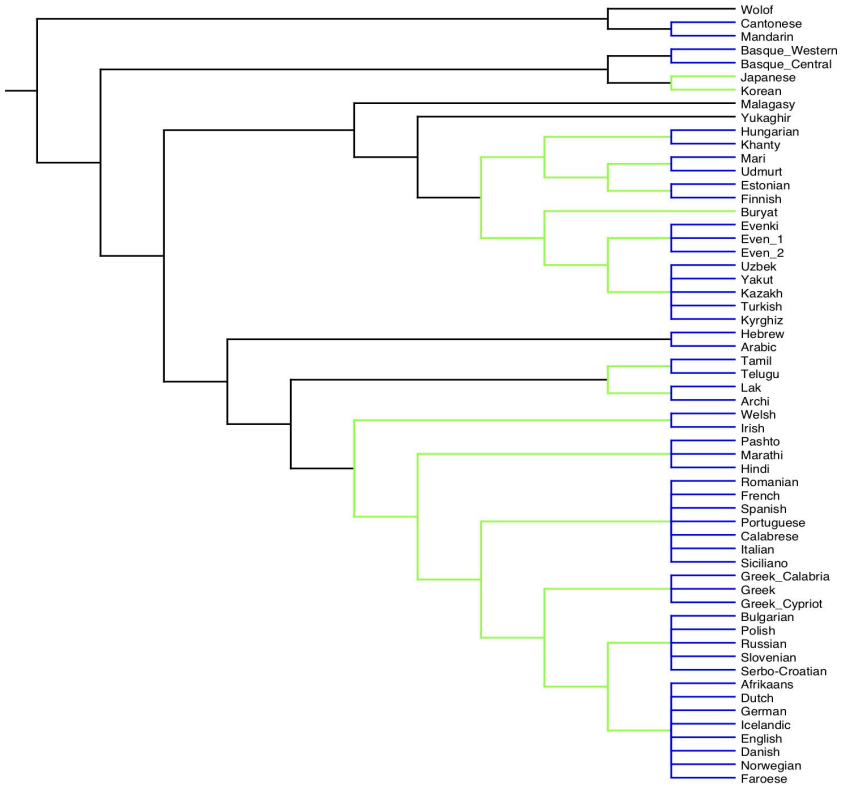

Fig 1. A phylogenetic UPGMA tree from syntactic distances. The branches which pass the significance test are in green.

Anderson S. and D. Lightfoot. (2002). *The language organ. Linguistics as cognitive physiology*. CUP. **Bortolussi L. et al. (2011).** How many possible languages are there? In G. Bel-Enguix, V. Dahl, and M. D. Jiménez-López (eds.), *Biology, computation and linguistics*, 168–179. M.D. IOS Press. **Ceolin A et al. (2020)** Formal Syntax and Deep History. *Front. Psychol.* 11:488871. **Jinam T. A. et al. (2015).** Unique characteristics of the Ainu population in Northern Japan. *Journal of Human Genetics* 60 (10): 565-571. **Kessler B. and A. Lehtonen. (2006).** Multilateral comparison and significance testing of the Indo-Uralic question. In Forster P. and C. Renfrew (eds.), *Phylogenetic methods and the prehistory of languages*, 33-42. McDonald Institute for Archaeological Research. **Lazaridis I. et al. (2016).** Genomic insights into the origin of farming in the ancient Near East. *Nature* 536 (7617): 419-424. **Longobardi G. and C. Guardiano. (2009).** Evidence for syntax as a signal of historical relatedness. *Lingua* 119 (11): 1679–1706. **Oswalt, R. L. (1970).** The detection of remote linguistic relationships. *Computer Studies in the Humanities and Verbal Behavior* 3: 117-129. **Reich D. (2018).** *Who we are and how we got here: Ancient DNA and the new science of the human past*. OUP. **Ringe D. (1992).** On calculating the factor of chance in language comparison. *Trans. Am. Phil. Soc.* 82 (1): 1-110. **Santos P. et al. (2020).** More Rule than Exception: Parallel Evidence of Ancient Migrations in Grammars and Genomes of Finno-Ugric Speakers. *Genes* 11(12):1491.