

Using Distributional Phylogenetic Modeling to Analyze Language Contact in Eastern Turkey

Nora Muheim
nora.muheim@helsinki.fi
General Linguistics

Research Aim

Traditionally, it has been assumed that Zazaki (diml1238, Northwest Iranian) is archaic, preserving old features of Iranian, e.g., the *ezāfe* linker is fully inflected as it was the case in Old Persian (Yakubovich 2020: 103). Recently, in the GramAdapt project, it has been shown that Zazaki has converged in the domain of nominal possession with Turkish (nucl1301, Common Turkic) (Ahola et al. 2023). To reconcile these two opposing findings, I approach the question of language contact in Eastern Turkey between the two Iranian languages, Zazaki and Northern Kurdish (nort2641), and Turkish by applying Bayesian statistical methods from a phylogeographical viewpoint, focusing again on nominal possession.

Methodology

The sampling method is adapted from Di Garbo & Napoleão de Souza (2023). There is at least one Focus Language (Zazaki and Northern Kurdish), a Neighbor (Turkish), which is in contact with the Focus, and a Benchmark, which is, in their work, related to the Focus but not geographically adjacent to either the Focus or the Neighbor. The Benchmark functions as a proxy in the absence of historical data. Here, the approach is adapted by creating a Benchmark based on *Ancestral State Reconstruction*. The questionnaire targets unmodified **possessive noun phrases**, as illustrated in the example below, taking the questionnaire from GramAdapt (cf. e.g., Ahola et al. 2023).

Zazaki (Todd 2008: 93)

pize- y mIn
stomach- EZ.SG.M me:OBL
'my stomach'

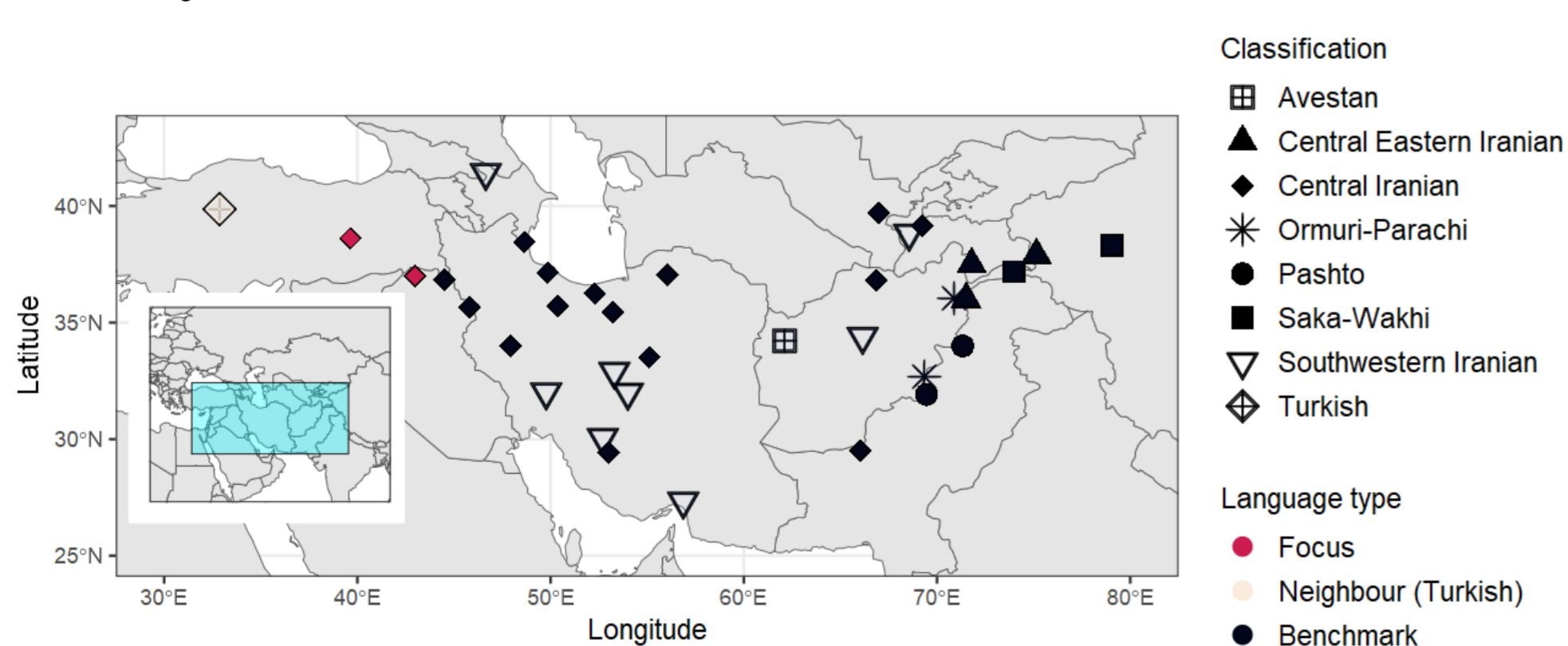


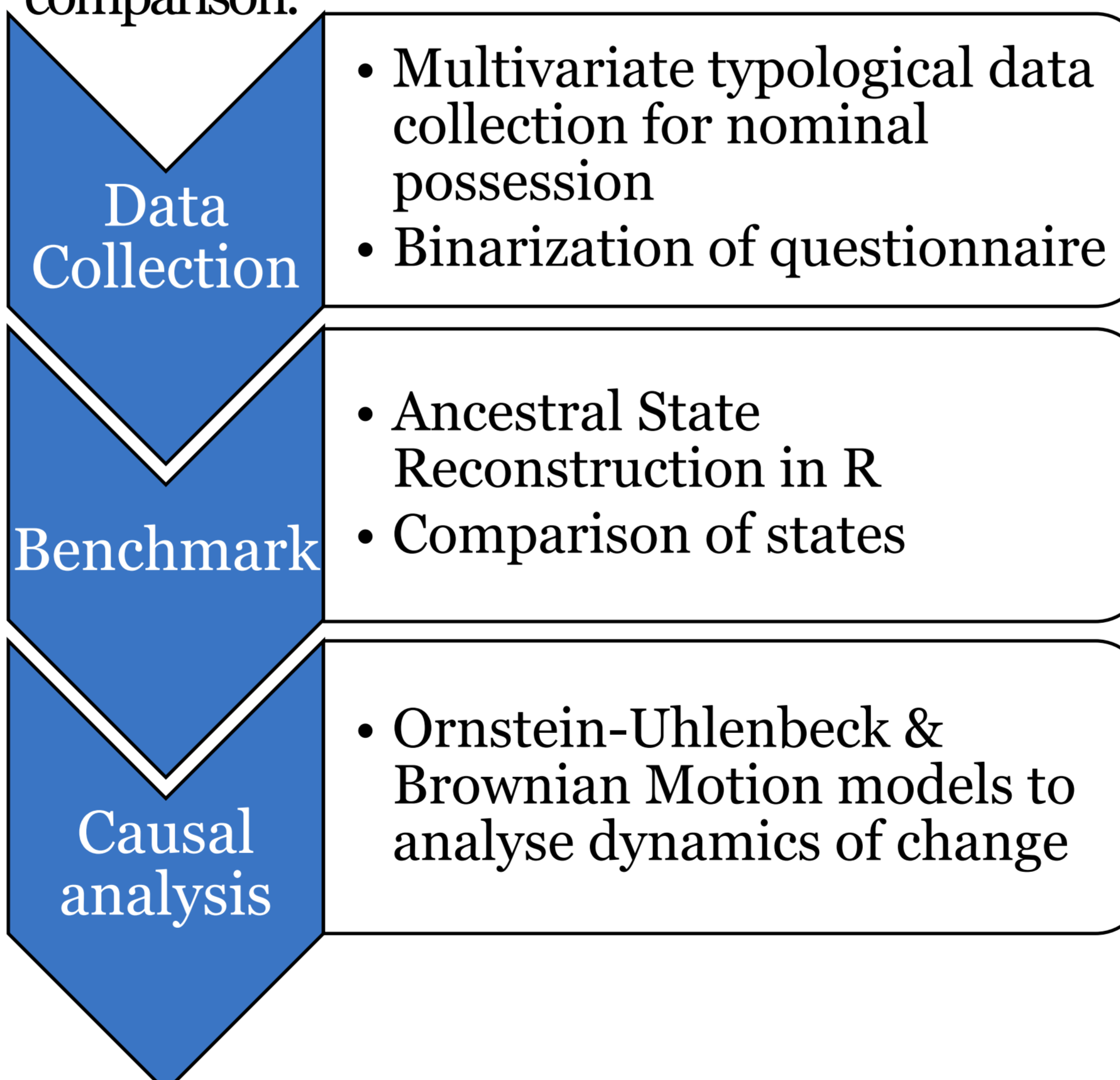
Figure 1: Distribution and classification of all languages in the sample, colored according to their function in the sample.

The questionnaire targets morphosyntactic features,

namely the locus of marking, conditioning factors, boundedness of the marking, the linear order of elements, and the position of the markers.

Workflow

The ancestral state reconstruction in step 2 is done in R Studio based on the code used in Sinnemäki & Ahola (2023). The AncThresh method from the phytools package (Revell 2024) was used with sample size 1000, burn-in 200,000, and 2,200,000 iterations with Brownian Motion. The results for one feature are illustrated in Figure 2. A second reconstruction with the Ornstein-Uhlenbeck model is planned for comparison.



In a future step, the focus will be on the different dynamics on the tree edges within the tree of Iranian languages in the Bayesian analysis (cf. Cathcart et al. 2023). To do so, Brownian motion and Ornstein-Uhlenbeck models –both simple and multivariate – are used (cf. Beaulieu et al. 2012). The models are implemented in Rstan. Brownian motion allows us to calculate the amount of shared history between different nodes. The Ornstein-Uhlenbeck model can tell us more about the length of divergence between two nodes. The difference between univariate and multivariate models means that the univariate ones only allow one unified rate of evolution, while the multivariate ones with two regimes allow two different rates (heterotachy).

Interim Results

Figure 2 shows the first results for one feature of the questionnaire, namely the presence or absence of head-marked pronominal possessive noun phrases.

As illustrated in Figure 2, this type of construction is predominantly present in the Avestan, Pashto, Saka-Wakhi, and Central Eastern Iranian Branches, but there are exceptions, e.g., Gilaki, Sorkhei Aftari, Sogdian, and Bactrian of Central Iranian.

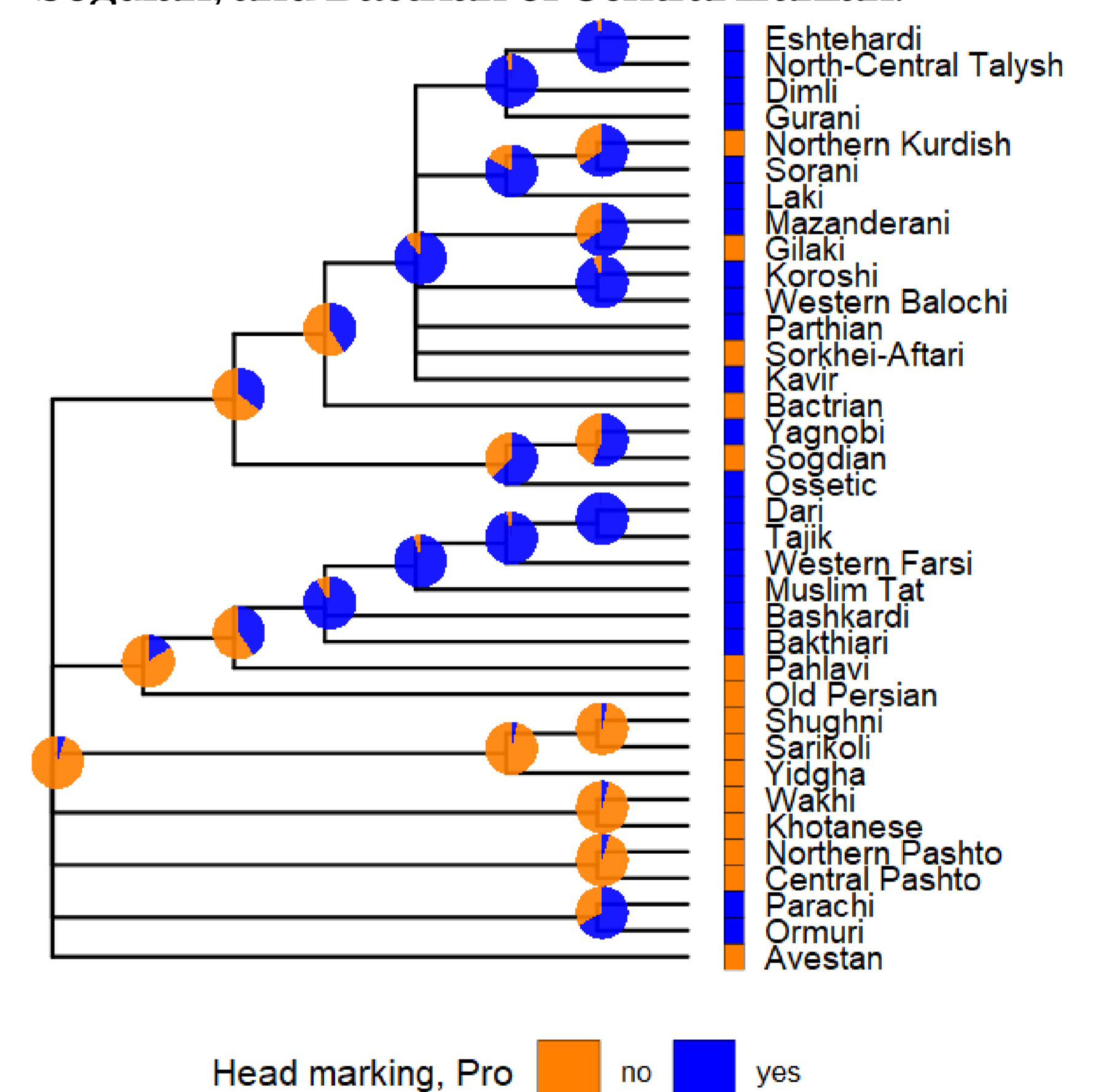


Figure 2: Bayesian ancestral state reconstruction for head marking of pronominal possessive noun phrases. Blue indicates the presence of pronominal head marking, while orange indicates the absence. The root of the tree is Proto-Iranian.

Zooming in on the Focus languages, a mixed picture emerges: the construction is present in Turkish and Zazaki, but absent in Northern Kurdish. The ancestral state reconstruction yields a 90.5% likelihood for the presence of head marking in pronominal possessive NPs in Northwest Iranian; hence, Zazaki aligns with this. On the other hand, the reconstruction shifts when going up the tree, resulting in a probability of 64.6% for the absence of this construction, so that Northern Kurdish aligns better with this result. It may be concluded from this that Northern Kurdish is more archaic with respect to this feature than Zazaki. Zazaki has the same construction as in Turkish, so it may have changed due to contact with it but on the other hand, Zazaki follows a wider trend in Northwestern Iranian.

References & Data

https://github.com/Kulmkapp/Kulmkapp.github.io/blob/main/sources_Salos2025.pdf (Sources)

<https://shorturl.at/tudNb> (Data)

