Otto-Friedrich-Universität Bamberg
Fakultät Geistes- und Kulturwissenschaften
Lehrstuhl für Allgemeine Sprachwissenschaft
Erweiterungsbereich, 8 ECTS
Prof. Dr. Geoffrey Haig
Sommersemester 2021

How language changes with special reference to syntax

Bayesian phylogenetic tree of modern Iranian doculects

Alexandru Craevschi

Contents

1 Language classification and the Iranian language family

1.1 A brief history and methods of linguistic classification

Language classification is one of the classical tasks and oldest aims of historical linguistics (Campbell & Poser, 2008; Campbell, 2013). It should be noted that historical linguistics is primarily interested in establishing "genetic relatedness among languages and language families" (Luraghi, 2017, p. 95), as opposed to linguistic typology which classifies languages based on their structural properties. An accurate and reliable language classification is an extremely valuable product of work not only for linguists but also provides evidence and insights for archaeologists, historians and anthropologists (Heggarty, 2014; Fortson IV, 2009). One important distinction that needs to be drawn are two levels of classification: internal and external classification (Starostin, 2009). Starostin (2009, p. 158) summarizes this distinction well in his review of Campbell and Poser (2008):

> Second, for the purposes of this particular volume, "classifying" languages is primarily understood as "demonstrating genetic relationship" between two or more languages or language groups (we may call this external classification), rather than estimating the degree of relationship between languages in an already well-established language group (internal classification). This is an important point, because in comparative linguistics, issues of internal classification are just as frequent and hotly debated (sometimes even more so) as those of external classification.

Joseph Greenberg considered that linguistic sub-grouping was even more complicated than the establishment of the main families (Winston, 1966, p. 166–167).

The main method used for both kinds of classification still remains the comparative method (Weiss, 2014). It remains the main tool of historical linguists since the XIX century, when Grimm formulated what is considered to be the first sound correspondences between the Proto-Germanic languages and Sanskrit (Weiss, 2014, p. 128). The comparative method establishes regular sound correspondences between the phonemes of a set of words of two or more languages that are hypothesized to be related. The same method

is used for linguistic reconstruction of ancient languages, also called proto-languages (Fox, 1995).

1.2 Classification of the Iranian language family

The Iranian language family is a branch[1] of a larger Indo-Iranian family (Windfuhr, 2009b), which in turn represents one of the major branches of the Indo-European language family (Fortson IV, 2009, p. 10). Despite the fact that the Indo-European language family is one of the best described and studied language families in the world, there still remain some issues in its internal classification.

Windfuhr (2009a, p. 21) uses more conventional East/West distinction when speaking of phonological innovations in the Iranian language family. Eastern and Western Iranian are the two hypothesized major branches of the Iranian language family. That is, languages that are classified as Eastern Iranian languages have a common ancestor, the Proto-Eastern-Iranian (PEI), and the same is true for Western Iranian languages, with their common ancestor conventionally being called the Proto-Western-Iranian (PWI). This view is challenged by a few researchers, among them Sims-Williams (1996); Korn (2017, 2016) and Cathcart (2015). Sims-Williams (1996, p. 651), for instance, concludes the following:

> It does not seem possible to regard the Eastern Iranian group as a whole – even disregarding Parachi and Ormuri – as a genetic grouping. Such a conception would imply the existence of an ancestral "proto-Eastern Iranian" (...); but if one reconstructs "proto-Eastern Iranian" in such a way as to account for all the features of the group, it proves to be identical to the "common Iranian" reconstructible as the ancestor of the whole Iranian family.

The similarities between the languages normally classified as the Eastern Iranian languages is said to be due to language contact.

Using two different computational methods, Cathcart (2015) gets to a similar conclusion. One of the methods he used was character-based

[1] In the next chapter, phylogenetic methods used for language classification will be briefly introduced. A technical term used in phylogenetics for a branch or sub-family is *clade* (Dunn, 2014).

Bayesian phylogeny.[2] First, he used a set of typological features of several Iranian languages and the tree (Cathcart, 2015, p. 50) that was produced shows support for East/West division. Conversely, when the phylogeny inferred the tree based on 200 Swadesh wordlist, the Eastern Iranian clade was not produced (Cathcart, 2015, p. 52). There is an explanation for why typological data yields this result and I will provide an overview of it in Section 2, but for now the short argument is all about language contact and the kind of features one looks at. Although somewhat counter-intuitive, structural features are mostly faster in their rates of change than the basic vocabulary (Greenhill et al., 2017). Wordlists of basic vocabulary showed to be a more reliable kind of data for language phylogenies (Greenhill, Heggarty, & Gray, 2020, p. 232–239).

2 Methods and data

2.1 Phylogenetic methods in linguistics

The last 20 years in historical linguistics were marked by importing various methods from evolutionary biology (Gray et al., 2007; List, 2014; Newberry et al., 2017). The most widely used method taken from evolutionary biology has undoubtedly been the inference of phylogenetic trees (Dunn, 2014; Greenhill et al., 2020; Goldstein, 2020): "Phylogenetic trees model linguistic descent. More specifically, they are hypotheses about the order of lineage-splitting events from an often unobservable common ancestor to a set of observable descendants" (Goldstein, 2020, p. 110). There exists a variety of algorithms that allow the inference of phylogenetic trees of languages. One of the most widely used ones nowadays is the family of methods called *character-based models of change* (Dunn, 2014, p. 196). Initially, when this kind of models were used in evolutionary biology, every character represented by an alphabet, for example, "the nucleotide (A, C, T, or G) that appears in a particular location within a gene, the number of legs (any positive integer), or whether the organism has hair (a Boolean variable)".

In case of language phylogenies, every character is represented by an

[2] In the next chapter I will discuss the intricacies of the model Cathcart (2015) used and point to some of the disadvantages of the substitution model he selected.

item from a borrowing resistant wordlist, such as Swadesh 100- or 200-items wordlist or Leipzig-Jakarta wordlist (Tadmor et al., 2010). The selected wordlist is then assembled for all the languages in question. After that, the comparative method is used to establish which of the words of two or more languages that express the same concept[3] are *cognates*, i.e. which words of two or more languages are genetically related and have a single ancestor word in their proto-language. The basic wordlists are used to minimize the problem of lexicon borrowing. When words are actively borrowed between two languages, one can mistakenly establish their genetic closeness. In this particular study I will use a variety of 100 wordlist called Leipzig-Jakarta wordlist (Tadmor et al., 2010). This wordlist is based on the Loanword Typology project which had as one of its aims, formulating empirically supported list of 100 items that are rarely borrowed to avoid the distortion caused by loanwords.

As soon as all the items were annotated and grouped into cognate sets,[4] one has to select a particular method of inferring a phylogeny. The most widely used method nowadays is Bayesian estimation of the posterior distribution of the parameter value (Goldstein, 2020, p. 154–156). In case of phylogenetic inference, our aim is to sample the posterior distribution of the parameter value. The parameter values are sampled in proportion to the likelihood of a set of parameters $\mathbf{X}$ that describe the tree, given the current data. Likelihood is "the relative number of ways that a value p can produce the data" (McElreath, 2020, p. 27). The set $\mathbf{X}$ includes parameters such as branch length, rate of variation and others.[5] Since the number of possible parameter values and their combinations is infinite, the stochastic algorithm called Marcov Chain Monte Carlo (MCMC) is used. MCMC spans the mathematical space and samples the parameter values. Every value is sampled proportionally to its representation in the posterior

[3]Note that only full cognates were used in this study. Sometimes partial cognates are also used. If two words have a common ancestor word but one of the languages has a different meaning for the word nowadays, then these two words are counted as partial cognates.

[4]In practice researchers code cognates by assigning them the same letter or the same number.

[5]The exact number of parameters in $\mathbf{X}$ varies and depends on the particular substitution and clock models.

distribution. The ultimate product of this procedure is not a single tree but rather a distribution of likely parameter values and trees.

An important modelling question is the selection of substitution model. Substitution model is a mathematical description of the transitions that may occur in the character (cognate) state along a tree. To exemplify it, a very popular model, stochastic Dollo, incorporates assumption that cognates can only be acquired one but can be lost several times (Bouckaert & Robbeets, 2017). The unfortunate problem with the stochastic Dollo model is that in practice it tends to provide a poor fit when there are coding errors or more or less extensive borrowing events.

Turning back to the Iranian language tree presented in Cathcart (2015, p. 49), we observe that Cathcart used precisely the stochastic Dollo model which is not very suitable in case of the Iranian languages due to the problems of language contact and its effect on the fit of the model. Most of the Iranian doculects in my sample (see the next Section) are minority languages and thus are heavily influenced by other major languages in the area (Haig & Khan, 2018). Moreover, many of the languages borrow lexicon from other Iranian languages which further introduces distortions in the tree model of language evolution. Because of that, the tree thinking imported from evolutionary biology (Baum & Smith, 2012) should be used with a certain degree of caution and awareness about the problem of horizontal transmission in language. For this reason, I will also present the δ-score and Q-residual (Kolipakam et al., 2018), two statistic of how "tree-like" the data in question is.

To avoid the problems of stochastic Dollo model described above, I have tried two other models as described in Bouckaert and Robbeets (2017) and implemented in BEAST (Drummond & Bouckaert, 2015), namely binary Continuous time Marcov chain (CTMC) and binary covarion models. The latter is especially popular because of the good fit it provides (Greenhill et al., 2020, p. 231). I also used a combination of the lexical and typological data to produce a single consensus tree, which is slightly closer to the way that classical linguistic classification works, as linguists normally take into account multiple factors when classifying languages, rather than only lexicon or only typology, see chapter 7 from Campbell and Poser (2008,

p. 162–223) for more details. Since we do not expect the lexicon and typological features to evolve in the same way, I used Bayesian Stochastic Variable Selection (BSVS) model with asymmetric transition probabilities. Asymmetric transition probabilities mean that the transition of character from the state A to the state B and vice versa in, say word order, are not equal.[6] The BSVS model is also implemented in BEAST (Drummond & Bouckaert, 2015).

Finally, we get to the question of clock model. Clock is essentially a measure of evolutionary changes that are used to estimate the branch lengths of a tree. Branch lengths show the time when the divergence between two taxa (doculects) occur. Relaxed clock models were used to estimate when a proto-language was spoken, as in the case of Proto-Indo-European (Bouckaert, Remco R., Lemey, Philippe, Dunn, Michael, Greenhill, Simon J., Alekseyenko, Alexander V., Drummond, Alexei J., Gray, Russell D., Suchard, Marc A. & Atkinson, Quentin, 2012; Chang et al., 2015), for instance. Based on the recommendations provided in Drummond and Bouckaert (2015, p. 145–148), I first ran a model with random clock for both lexical and typological data. After inspecting the output with the Tracer, the software that comes along with the BEAST software, it was decided to use the relaxed clock for lexical data and strict clock for typological data. This means that rates of change along the tree branches are allowed to differ. I.e. albeit the number of evolutionary changes along the two different branches of the tree are the same, the time frame in which they happened is allowed to be different. The models presented in the Section 3 were run three times to ensure good mixing of MCMC and to yield the necessary number of ESS (Drummond & Bouckaert, 2015, p. 140–145). The results shown in Section 3 are thus the combined result of the three runs. In all the cases between 25% and 35% of burn-in were discarded from the analysis. These routines are required due to the stochastic nature of MCMC algorithm, see McElreath (2020, p. 263–296) for more details.

[6] Suppose a proto-language had SOV word order, while some of its descendants now have SVO word because of the areal pressure. Thus, SOV→SVO transition is more likely than SVO→SOV transition in this case, since the latter would require two changes: from proto SOV to SVO and then back to SOV.

2.2 Data: Sources and processing routines

I used 23 Iranian languages and dialects for this study. I will refer to them conventionally "doculects" to avoid the sloppiness in the definition of notions "language" and "dialect". The languages used for the study and their geographical distribution is shown on the Figure 1. The various sources used in the study to collect lexical and typological data are presented in two Excel files that can be found in the GitHub repository.[7]

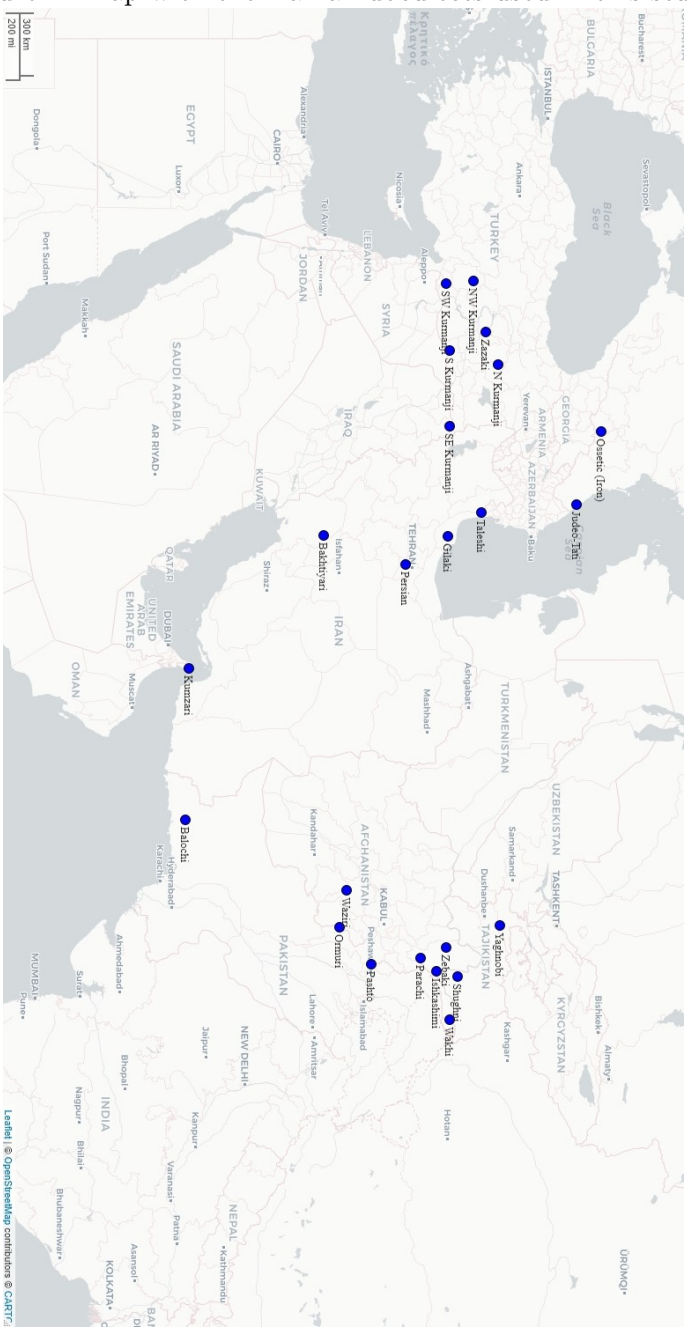
With respect to the lexical data, I decided to use the Leipzig-Jakarta 100-items wordlist (Tadmor et al., 2010) for this study as it should reduce the impact of the unmarked loanwords to the minimum. For some of the doculects, like Zebaki for instance, the data available was very scarce but nonetheless enough to group it correctly with its undisputed sister Ishkashmi.[8]

The data were first collected and organized in a format suitable to analyse it with the help of LingPy library (List & Forkel, 2021). I first tokenized the words and then aligned the tokens. The output from these procedures was then fed into a function that automatically detected the cognates and grouped them in cognate sets. The output file produced by the LingPy was then manually inspected and corrected. All the errors in the cognate coding are thus mine. The data, code, LingPy's log files and my correction log can all be checked by accessing the GitHub repository. In the data that were first used for automated cognate detection and afterwards in the correction log, one can observe doculect "Luri" which is a sister variety of Bakhtiyari. I decided to discard Luri from the phylogenetic inference, as the data I initially used were found in Cathcart (2015) but I could not find the source of the lexical data that would allow me to verify the existing entries from Cathcart's thesis. The corrected lexical data were subsequently used to calculate the Hamming distances between the languages. These distances allow to construct a NeighborNet network in SplitsTree (Huson & Bryant, 2006). One of the most valuable metrics that NeighborNet subsequently allows to compute is δ-score and Q-residuals which are metrics of how tree-like the data is.

[7] `https://github.com/acraevschi/Iranian_phylogeny`

[8] To my knowledge there are no argues about Ishkashmi and Zebaki being two varieties or closely related dialects of the same language.

Figure 1: Map with the Iranian doculects used in this study



The 9 typological features used in this study were carefully selected. As I have mentioned in Section 1.2, typological features have been shown to change mostly at faster rates than the basic lexicon from 100- or 200-items wordlists, such as Leipzig-Jakarta wordlist. The same study by Greenhill et al. (2017) also showed that there are some features that are as resistant to change as the basic vocabulary is. Thus, in this study I decided to use only the typological features that change slowly, to make them reliable for language classification and to minimize the confounder effect of language contact. The features for this study were selected from the previously mentioned paper by Greenhill et al. (2017) and also from a paper with similar research question by Dediu and Levinson (2012). These papers used two different phylogenetic methods to estimate rates of change of structural features in different language families. The features were then found in the World Atlas of Language Structures (Dryer & Haspelmath, 2013) and coded in the same way for consistency for the languages in question. The full list of features and values can be seen in the table in GitHub repository.

3 Analysis results

3.1 Phylogenetic tree of modern Iranian doculects

All the relevant visualizations of the trees can be found in Appendix 1.

Eastern and Western clades are produced in 60% when using only Leipzig-Jakarta wordlist and CTMC as the substitution model. Ossetic[9] is not included in either of the clades and forms a separate branch of its now. That is the truth for 3 out of the 4 maximum clade credibility trees represented in this study. In the tree represented on the Figure 5 Ossetic is included in the Western Iranian group but even then there is a lot of uncertainty about it being a separate member of Western Iranian, as can be seen in the density tree. The problem with its classification is though easily explainable:

> Ossetic, like its Alanic predecessor, has for millennia been separated from the sister languages of Central Asia, being spoken in non-Iranian surroundings. It has developed certain characteristic peculiarities, in

[9]Since only one variety of Ossetic, Ossetic Iron, is represented in the study, from now on I will use the term "Ossetic" to refer to this variety

> part due to the influence of adjacent languages (Turkic, Caucasic). This applies to vocabulary as well as phonetic and grammatical structure. As regards lexical borrowing, the influence of Turkic languages seems to have been particularly strong. (Thordarson, 2009)

A seeming error of the phylogeny occurs in the classification of Talyshi, as it should form part of the Western Iranian group (Asatrian & Borjian, 2005, p. 51). Unfortunately, no Tatic languages were included in the study, as Talyshi should form a separate clade with them. Judeo-Tati is different from the other Tatic languages and should not be confused with them. Overall, as noted by Stilo (1981), Tati is not a unified set of doculects and it is difficult to discern genetic and areal factors. Nonetheless, in both genetic and areal influence, Talyshi should be grouped with other Western Iranian languages. This might be a peculiarity of items found Leipzig-Jakarta list or due to coding errors.

Another interesting case is the classification of Parachi and Ormuri, as these two languages have long posed difficulties for linguists. I have already touched this problem superficially by citing Sims-Williams (1996, p. 651) in Section 1.2. A more detailed overview and new contribution to the question of classification of Ormuri and Parachi can be found in Trofimov (2018). Trofimov establishes sound correspondences on the material of 110-item Swadesh wordlist for Parachi and Ormuri and compares them with Western and Eastern Iranian innovations.[10] He concludes that Parachi and Ormuri are related to other Eastern Iranian languages and should not be classified as Western Iranian languages, despite there being some discussions of this idea (Trofimov, 2018, p. 281). Trofimov does not go as far to determine the exact subgrouping of Parachi and Ormuri, though. In case of the trees produced by the models I employed, Parachi and Ormuri are grouped either as: i) Parachi and Ormuri being Pamir languages; or ii) Parachi being closely related to Pashto and Ormuri included in the Pamir languages. Importantly, in either case, the maximum clade credibility trees do not produce a branch that would prove Parachi and Ormuri being directly related sisters. Note that Pamir languages are not a unified language family,[11] I rather use this

[10] Note the drawback that Eastern/Western clades seem to be assumed to exist by Trofimov.

[11] "...the term "Pamir languages" is based on their geographical position rather than on

notion as a covert term for the Eastern Iranian languages that do not group together with Pashto and its varieties from a more traditional point of view. By more traditional, I have in mind a tree from Glottolog (Hammarström et al., 2021), for example.

The further division of the Eastern and Western clades into yet another two branches was also previously proposed in the literature. The Eastern clade separates into South Eastern Iranian and North Eastern Iranian (Novák, 2013, p. 23–59), the same names are often used for the two clades in the Western Iranian branch (Windfuhr, 2009a, p. 13–15). The North Western Iranian languages are represented by Zazaki and the varieties of Kurdish, while the rest forms South Western group. Notably, despite the problems with this kind of classification (Novák, 2013, p. 32), this is what most of the trees in the sample yield. Only the covarion substitution model based on the lexical data only does not produce this subdivision of the Eastern group because no single Proto-Eastern Iranian is produced in the first place. In turn, all the models with different kinds of data reliably produce this division in the Western clade. Thus, the trees reflect the current state of knowledge about the internal subgrouping of Western and Eastern Iranian branches. The Western's one is relatively well established and the Eastern one contains a lot of uncertainty and this is reflected in the posterior probabilities of the Northern/Southern division in the trees.

According to the models' estimates, the southern branch of the Eastern Iranian is the oldest one with two closely related doculects Pashto and Waziri being its undeniable representatives, while Parachi seems to be somewhere in the middle between the South and North Eastern Iranian.

To some extent this is a good signal that the phylogenies adequately comply with the previous findings established by the comparative method as applied by different specialists. The latter initially was thought and meant to group languages in a tree-like manner (François, 2014, p. 162) but it was then continuously questioned as whether tree models are appropriate because of the horizontal transmission that often occurs. This matter will be discussed in more details in the following subsection.

---

their genetic proximity, and they have also been called a "Sprachbund" (linguistic area), which seems to be more appropriate."

In all of the cases, the Eastern Iranian branch is estimated to be older than the Western Iranian branch. Even though I wanted to focus on the topology of the Iranian languages, this observation provides additional evidence for the adequacy of phylogenies for glottochronology.[12] It is believed that the Eastern branch is more ancient based on indirect evidence: the oldest attested Eastern Iranian language, Avestan, is considerably more archaic than the oldest attested Western Iranian language – Old Iranian (Novák, 2013, p. 9). Even if we accept that Ossetic is a Western Iranian language (which is highly doubtful) as it is the case in the tree from the Figure 5, the Western clade is still slightly younger than the Eastern Iranian clade.

As I have previously pointed out, there is much more certainty in the classification of the Western Iranian subgroup. But even there we have some open questions. For instance, there exist some doubts as to whether Zazaki and Kurdish (Kurmanji in our case) are immediate sisters or not. Their verbal morphology and especially irregular verb forms are strikingly different.[13] In case of the phylogenies introduced here, Kurdish and Zazaki are grouped as sister languages but the models suggest that they separated a long time ago, soon after the diversification of Proto-Western Iranian into the Southern and the Northern clades. The languages are well identified for both subgroups, apart from Gilaki which should be attributed to the North Western Iranian groups (Stilo, 2012). If one examines carefully the density trees, it is easy to observe that there is a lot of uncertainty in the classification of Gilaki.

An interesting group to be found in the Western Iranian clade is that of Kumzari and Balochi. In all of the maximum clade credibility trees, Kumzari and Balochi are said to form a separate clade of the Western Iranian languages. But the posterior probability of this clade is always between 50% and 60%, i.e. this group is non-existent in almost half of the trees produced by the analysis. If we consider Balochi's migrations,[14] it looks possible that Kumzari and Balochi had a common ancestor. Nevertheless, this hypothesis

[12]Glottochronology is concerned with the age of individual languages and language families.

[13]I learned this from a personal discussion with Geoffrey Haig.

[14]These migrations have led to formation of several Balochi varieties that are quite different from one another. The variety used in the study is Southern Balochi.

needs to be tested by well-established methods, such as the comparative method.

3.2 Wave and tree model of language diversification

The difficulties and errors that I have touched upon in the previous section are not particularly new for the fields of language classification and historical linguistics. Several times "Sprachbund" or "areal influence" have been mentioned to account for the uncertainty present in the Eastern Iranian clade. The areal influence, often also called "horizontal transmission" when discussing language classification, is a phenomenon when languages that need to be classified are passing linguistic material (e.g. lexicon) not only from proto-language to its descendants but rather transmission also occurs between two or more simultaneously existing languages. One shall keep in mind that methods used in this paper are imported from evolutionary biology and population genetics, and gene exchange between different taxa is rarely the case.

Thus, one of the major criticisms of the phylogenetic methods in linguistics has been the fact that the method is unable to account for horizontal transmission. In a series of simulations, Greenhill et al. (2009) showed that "realistic levels of reticulation between cultures do not invalidate a phylogenetic approach to cultural and linguistic evolution". Despite that, there are a few specific cases when a phylogeny or, more generally, the classification of correct groups and sub-groups is extremely difficult. One of the examples is introduced in Bowern (2013, p. 427):

> The second case is where languages have gradually diverged *in situ* and non-overlapping isoglosses remain from the old dialect area ... This situation is similar to that described by Ross' (1997) linkage model, which he characterizes as *"the (usually gradual) geographic spread of a group of speakers"* (Ross, 1997: 212) ... Networks of this type, however, are messy because of divergence processes; that is, it is not contact between related languages that directly produces ambiguities in discrete subgrouping, but rather conflicting language split.

For such cases, wave model of language diversification was developed (Kalyan & François, 2018; François, 2014). This model aims at integrating both phylogenetic and areal influence to account for divergence and con-

vergence of genetically related languages. The wave model, despite some success,[15] still stays behind in terms of its complexity and the number of possibilities it offers. For instance, Bayesian phylogenetics allows researcher to provide calibrations for language dating and even reconstruct the proto-language through phylogeny. That is, if some knowledge about a language community is provided by means of archaeology or there exist some written records, this knowledge can be easily taken into account by the modern phylogenetic techniques. This kind of modelling is yet to be implemented in case of the wave model.

Nevertheless, the wave model and linkage thinking should continue to develop as it potentially could solve the problem of the kind that the Iranian sub-grouping poses for the current phylogenetic methods. The final metrics that are important to inspect when dealing with phylogeny are δ-score and Q-residuals. The two metrics provide a descriptive statistic about the data used in the model and quantifies how tree-like the data is. A detailed explanation about the algorithm to compute δ and Q-residual scores is described in Gray et al. (2010, p. 3925). An example used in Gray et al. (2010, p. 3926) is that of Sranan, an English and Dutch lexified creole. Given the history and development of pidgins and creoles (Velupillai, 2015), they follow a very non-tree like model of diversification because of a multitude of factors and confounders that condition their structure and lexicon. (Gray et al., 2010) computed δ-score and Q-residuals for Sravan. It was included in a dataset along with 12 other Indo-European languages, among which we find Dutch and English. δ-score and Q-residuals for Sravan were 0.29 and 0.05 respectively (Gray et al., 2010, p. 3926).[16] The lower the value, the more tree-like the data is. δ-score and Q-residuals for the doculects in my dataset are presented in the Table 1.

It is obviously interesting to compare the tree-likeness of the Iranian language family with that of the other language families, both macro families and some smaller groups. The data in the Table 2 are taken from Kolipakam

[15] For practical demonstration of the wave model application, see Kalyan and François (2018).

[16] Both δ-score and Q-residuals are computed for every individual language and then summarized for the entire dataset. In case of Sravan and Indo-European data, the average δ-score was 0.23 and Q-residual = 0.03. I.e. Sravan's values were way above average.

Table 1: δ-scores and Q-residuals for Iranian doculects

Doculect	**δ-score**	**Q-residual**
Bakhtiyari	0.36125	0.0089118
Balochi	0.3893	0.0089613
Gilaki	0.34024	0.0078168
Ishkashimi	0.33592	0.00762
Judeo-Tati	0.34138	0.0078239
Kumzari	0.35742	0.0073342
Modern Persian	0.33448	0.0094043
NK	0.27129	0.0069489
NWK	0.27241	0.0072615
Ormuri	0.36225	0.0071904
Ossetic Iron	0.36427	0.0064467
Parachi	0.38523	0.0067105
Pashto	0.33159	0.0064243
SEK	0.28476	0.0069559
Shughni	0.35576	0.0095163
SK	0.26886	0.0055392
SWK	0.26365	0.006629
Taleshi	0.38436	0.0071723
Wakhi	0.39751	0.0091358
Waziri	0.33536	0.01032
Yaghnobi	0.38888	0.008679
Zazaki	0.38383	0.0091707
Zebaki	0.35507	0.0090604
Average	**0.3420**	**0.007871**

et al. (2018).

Table 2: δ-scores and Q-residuals for different language families

Family	**δ-score**	**Q-residual**
Austronesian	0.33	0.002
Indo-European	0.22	0.002
Polynesian	0.41	0.02
Dravidian	0.30	0.0069
Iranian	0.3420	0.007871

As we can see the Iranian language family is less tree-like than the Indo-

European one, of which it forms part, and even less tree-like than Dravidian. In case of Dravidian, its non-tree-likeness was explained by the multilingualism of the speakers of Dravidian. It is likely to be the case for Iranian languages too, as there are a few dominant languages that are very likely to serve as lingua franca for various and diverse linguistic communities. Take for instance Pashto and other Eastern Iranian languages. Pashto has around 32 million speakers, while all the other Eastern Iranian languages combined have less than a million speakers (Novák, 2013, p. 23). This creates an incentive for different language communities to learn Pashto as L2 or even completely switch to it by teaching and growing their children in Pashto environment.

4 Conclusion

In this paper I inferred the tree of the Iranian language family using Bayesian phylogenetic methods. I used two different site substitution models and two sets of data, lexical and typological, to see what the topology of the Iranian languages looks like. One model that used covarion substitution model and lexical data only has not produced the Eastern Iranian clade, while all the other models produce it. To further improve the inference one must either i) include some additional calibrations or ii) apply a different model, like one of the novel phylogenetic methods called approximate Bayesian computation (Hang Fan & Kubatko, 2011). There are not a lot of examples of applying approximate Bayesian computation for language phylogenies. We observed that Iranian language family is relatively non-tree-like. Nevertheless, the application of phylogenetic methods is useful in case we want to infer genealogical relatedness. Otherwise, it might be useful to apply the wave model, as exemplified in Kalyan and François (2018).

References

Asatrian, Garnik, & Borjian, Habib. (2005). Talish and the talishis (the state of research). *Iran & the Caucasus*, *9*, 43–72.

Baum, David A., & Smith, Stacey D. (2012). *Tree thinking: An introduction to phylogenetic biology.* Macmillan Learning.

Bouckaert, Remco R., & Robbeets, Martine. (2017). *Pseudo Dollo models for the evolution of binary characters along a tree.* (preprint) doi: http://dx.doi.org/10.1101/207571

Bouckaert, Remco R., Lemey, Philippe, Dunn, Michael, Greenhill, Simon J., Alekseyenko, Alexander V., Drummond, Alexei J., Gray, Russell D., Suchard, Marc A. & Atkinson, Quentin. (2012). Mapping the origins and expansion of the Indo-European language family. *Science*, *337*(6097), 957–960. doi: 10.1126/science.1219669

Bowern, Calire. (2013). Relatedness as a factor in language contact. *Journal of Language Contact*, 411–432. doi: 10.1163/19552629-00602010

Campbell, Lyle. (2013). *Historical linguistics: An introduction* (3rd ed.). Edinburgh University Press. Retrieved from `http://www.jstor.org/stable/10.3366/j.ctt1g0b5gq`

Campbell, Lyle, & Poser, William J. (2008). *Language classification: History and method.* Cambridge University Press. doi: https://doi.org/10.1017/CBO9780511486906

Cathcart, Chundra. (2015). *Iranian dialectology and dialectometry* (Doctoral dissertation, UC Berkeley). Retrieved from `https://escholarship.org/uc/item/5pv6f9g9`

Chang, Will, Cathcart, Chundra, Hall, David, & Garrett, Andrew. (2015). Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language*, *91*(1). doi: 10.1353/lan.2015.0007

Dediu, Dan, & Levinson, Stephen C. (2012). Abstract profiles of structural stability point to universal tendencies, family-specific factors, and ancient connections between languages. *PLoS One*, *7*(10). doi: 10.1371/journal.pone.0045198

Drummond, Alexei J., & Bouckaert, Remco R. (2015). *Bayesian evolutionary analysis with BEAST.* Cambridge University Press. doi:

10.1017/CBO9781139095112

Dryer, Matthew S., & Haspelmath, Martin (Eds.). (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology. Retrieved from `https://wals.info/`

Dunn, Michael. (2014). Language phylogenies. In Claire Bowern & Bethwyn Evans (Eds.), *The Routledge handbook of historical linguistics* (pp. 190–211). Routledge. doi: https://doi.org/10.4324/9781315794013

Fortson IV, Benjamin W. (2009). *Indo-European language and culture: An introduction, 2nd edition*. Wiley-Blackwell.

Fox, Anthony. (1995). *Linguistic reconstruction: An introduction to theory and method*. Oxford University Press.

François, Alexandre. (2014). Trees, waves and linkages: Models of language diversification. In Claire Bowern & Bethwyn Evans (Eds.), *The Routledge handbook of historical linguistics* (pp. 161–189). Routledge. doi: https://doi.org/10.4324/9781315794013

Goldstein, David. (2020). Indo-European phylogenetics with R: A tutorial introduction. *Indo-European Linguistics*, *8*(1), 110 - 180. doi: https://doi.org/10.1163/22125892-20201000

Gray, Russel D., Bryan, David, & Greenhill, Simon J. (2010). On the shape and fabric of human history. *Proceedings of the Royal Society B*, *365*, 3923–3933. doi: 10.1098/rstb.2010.0162

Gray, Russel D., Greenhill, Simon J., & Ross, Robert M. (2007). The pleasures and perils of darwinizing culture (with phylogenies). *Biological Theory*, 360–375. doi: https://doi.org/10.1162/biot.2007.2.4.360

Greenhill, Simon J., Currie, Thomas E., & Gray, Russell D. (2009). Does horizontal transmission invalidate cultural phylogenies? *Proceedings of the Royal Society B*, *276*, 2299–2306. doi: 10.1098/rspb.2008.1944

Greenhill, Simon J., Heggarty, Paul, & Gray, Russell D. (2020). Bayesian phylolinguistics. In Richard D. Janda, Brian D. Joseph, & Barbara S. Vance (Eds.), *The handbook of historical linguistics* (pp. 226–253). John Wiley and Sons. doi: https://doi.org/10.1002/9781118732168.ch11

Greenhill, Simon J., Wu, Chieh-Hsi, Hua, Xia, Dunn, Michael, Levinson, Stephen C., & Gray, Russell D. (2017). Evolutionary dynamics of

language systems. *Proceedings of the National Academy of Sciences*, *114*(42), 8822–8829. doi: 10.1073/pnas.1700388114

Haig, Geoffrey, & Khan, Geoffrey (Eds.). (2018). *The languages and linguistics of Western Asia: An areal perspective*. De Gruyter Mouton. doi: 10.1515/9783110421682

Hammarström, Harald, Forkel, Robert, Haspelmath, Martin, & Bank, Sebastian. (2021). *Glottolog 4.4*. Leipzig. Max Planck Institute for Evolutionary Anthropology. Retrieved from `https://glottolog.org/accessed2021-09-29` doi: 10.5281/zenodo.4761960

Hang Fan, Helen, & Kubatko, Laura S. (2011). Estimating species trees using approximate Bayesian computation. *Molecular Phylogenetics and Evolution*, *59*(2), 354–363.

Heggarty, Paul. (2014). Prehistory through language and archaeology. In Claire Bowern & Bethwyn Evans (Eds.), *The Routledge handbook of historical linguistics* (pp. 598–626). Routledge. doi: https://doi.org/10.4324/9781315794013

Huson, Daniel H., & Bryant, David. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, *23*, 254–267. doi: https://doi.org/10.1093/molbev/msj030

Kalyan, Siva, & François, Alexandre. (2018). Freeing the comparative method from the tree model: A framework for historical glottometry. In Kikusawa Ritsuko & Lawrence A. Reid (Eds.), *Let's talk about trees: Genetic relationships of languages and their phylogenic representation* (Vol. 98, pp. 59–89). Osaka: National Museum of Ethnology of Osaka.

Kolipakam, Vishnupriya, Jordan, Fiona M., Dunn, Michael, Greenhill, Simon J., Bouckaert, Remco R., Gray, Russell D., & Verkerk, Annemarie. (2018). A Bayesian phylogenetic study of the Dravidian language family. *Royal Society Open Science*, *5*(3), 171504. doi: 10.1098/rsos.171504

Korn, Agnes. (2016). A partial tree of Central Iranian: A new look at Iranian subphyla. *Indogermanische Forschungen*, 401–434. doi: https://doi.org/10.1515/if-2016-0021

Korn, Agnes. (2017). The evolution of Iranian. In Jader Klein, Brian Joseph, & Matthias Fritz (Eds.), *Handbook of comparative and historical Indo-*

European linguistics (pp. 609–624). De Gruyter Mouton. doi: https://doi.org/10.1515/9783110261288-038

List, Johann-Mattis. (2014). *Sequence comparison in historical linguistics.* Düsseldorf university press. doi: 10.1515/9783110720082

List, Johann-Mattis, & Forkel, Robert. (2021). *LingPy. a Python library for historical linguistics.* Max Planck Institute for Evolutionary Anthropology. Retrieved from `https://lingpy.org/` doi: https://zenodo.org/badge/latestdoi/5137/lingpy/lingpy

Luraghi, Silvia. (2017). Typology and historical linguistics. In Alexandra Y. Aikhenvald & R. M. W. Dixon (Eds.), *The Cambridge handbook of linguistic typology* (p. 95–123). Cambridge University Press. doi: 10.1017/9781316135716.004

McElreath, Richard. (2020). *Statistical rethinking: A Baeysian course with examples in R and Stan* (2nd ed.). Chapman & Hall/CRC.

Newberry, Mitchell G., Ahern, Christopher A., Clark, Robin, & Plotkin, Joshua B. (2017). Detecting evolutionary forces in language change. *Nature*, *551*, 223–226. doi: https://doi.org/10.1038/nature24455

Novák, Ľubomír. (2013). *Problem of archaism and innovation in the Eastern Iranian languages* (Unpublished doctoral dissertation). Charles University in Prague.

Sims-Williams, Nicholas. (1996). Eastern Iranian languages. In *Encyclopædia iranica* (Vol. 7, pp. 649–652).

Starostin, George. (2009). Review of: Language classification: History and method. by Lyle Campbell and William J. Poser. *Journal of Language Relationship*, *2*, 158–174.

Stilo, Donald L. (1981). The Tati Language Group in the Sociolinguistic Context of Northwestern Iran and Transcaucasia. *Iranian Studies*, *14*(3/4), 137–187. Retrieved from `https://www.jstor.org/stable/4310364`

Stilo, Donald L. (2012). *Gīlān.* Retrieved from `https://iranicaonline.org/articles/gilan-x`

Tadmor, Uri, Haspelmath, Martin, & Taylor, Bradley. (2010). Borrowability and the notion of basic vocabulary. *Diachronica*, 226–246. doi: 10.1075/dia.27.2.04tad

Thordarson, Fridrik. (2009). *Ossetic language i. history and description.* Retrieved from `https://www.iranicaonline.org/articles/ossetic`

Trofimov, Artem. (2018). On the place of Parachi and Ormuri among the Iranian languages according to the data of annotated Swadesh lists. *Journal of Language Relationship*, *16*(4), 277–292.

Velupillai, Viveka. (2015). *Pidgins, creoles and mixed languages: An introduction.* John Benjamins. doi: 10.1075/cll.48

Weiss, Michael. (2014). The comparative method. In Claire Bowern & Bethwyn Evans (Eds.), *The Routledge handbook of historical linguistics* (pp. 127–145). Routledge. doi: https://doi.org/10.4324/9781315794013

Windfuhr, Gernot. (2009a). Dialectology and topics. In *The Iranian languages* (pp. 5–43). Routledge.

Windfuhr, Gernot. (2009b). Introduction to the Iranian languages. In *The Iranian languages* (pp. 1–5). Routledge.

Winston, F. D. (1966). Greenberg's classification of African languages. *African Language Studies* 7, 160–171.

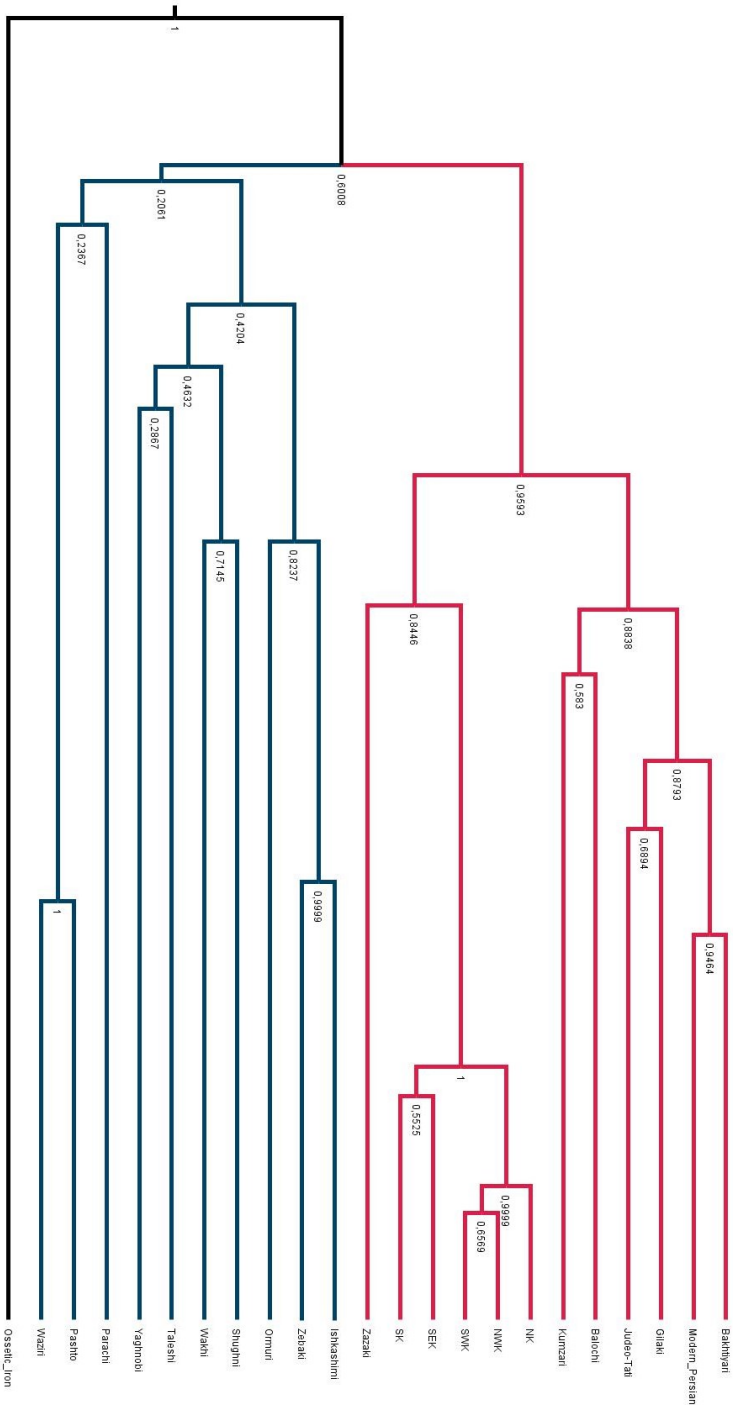
Appendices

Appendix 1

In the following Figures, maximum clade credibility trees of each of the models are presented. Red corresponds to West Iranian clade, whereas blue indicates East Iranian clade. Ossetic Iron is always marked with black as it is classified as West Iranian in just one one of the maximum clade credibility trees (Figure 5) and even in this tree the mean posterior probability of this clade is low. In Figure 4 East Iranian clade is not produced.

After each of the maximum clade credibility trees, one can also see density tree, visualized with the help of BEAST (Drummond & Bouckaert, 2015). This kind of trees visualizes the uncertainty about some of the groupings that are shown in the maximum clade credibility trees with numbers.

Figure 2: Lexical only phylogeny as produced by CTMC model



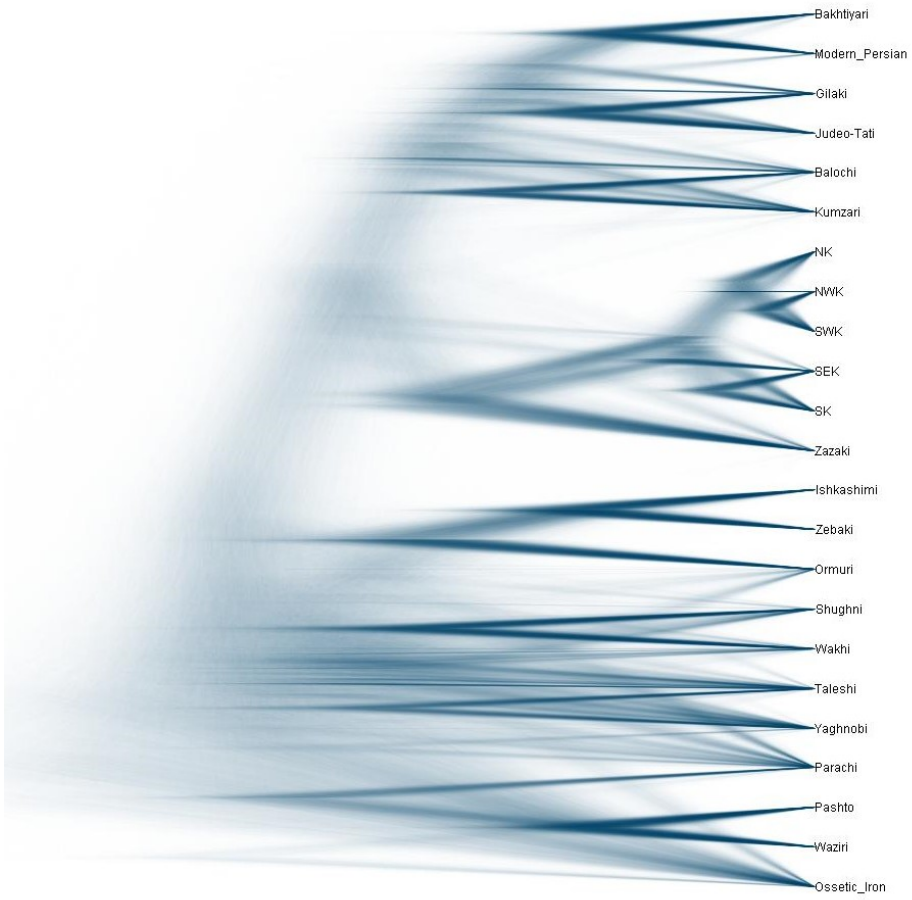
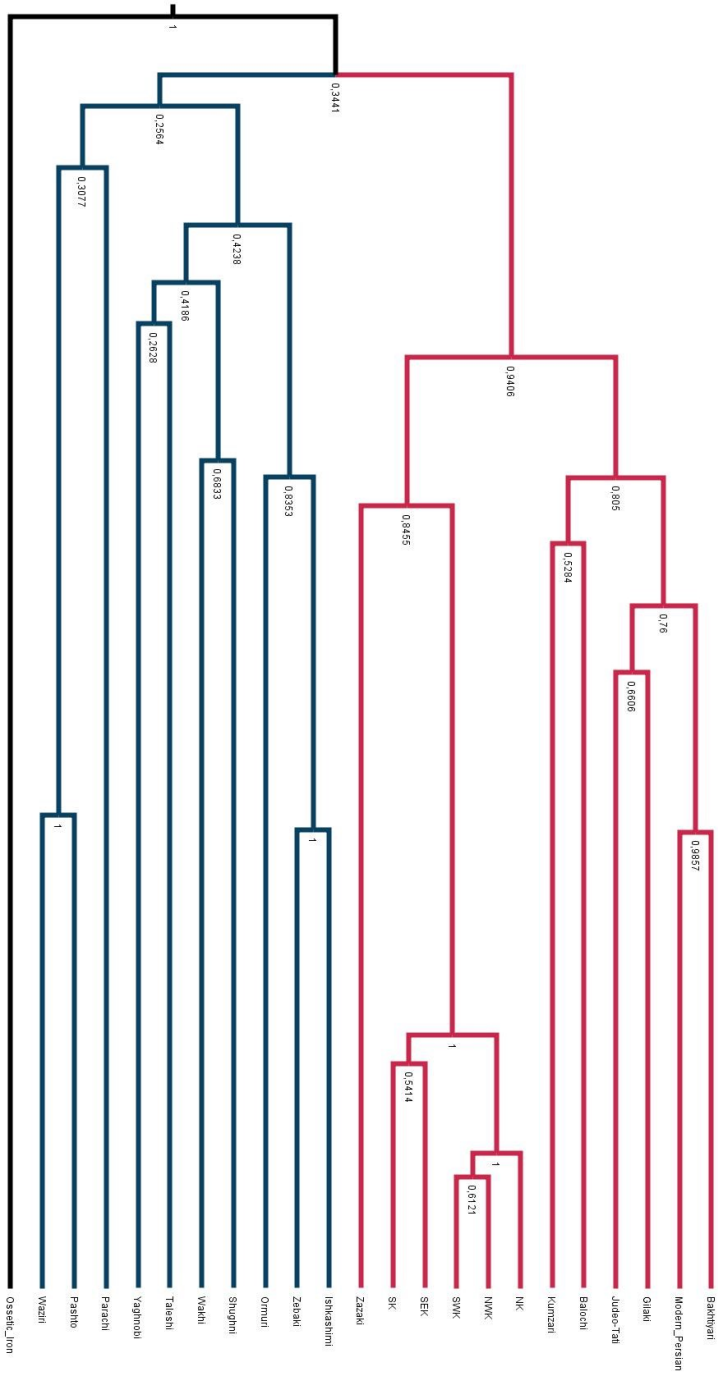

Bakhtiyari
Modern_Persian
Gilaki
Judeo-Tati
Balochi
Kumzari
NK
NWK
SWK
SEK
SK
Zazaki
Ishkashimi
Zebaki
Ormuri
Shughni
Wakhi
Taleshi
Yaghnobi
Parachi
Pashto
Waziri
Ossetic_Iron

Figure 3: Combined lexical and typological phylogeny as produced by CTMC and BSVS models for lexical and typological data



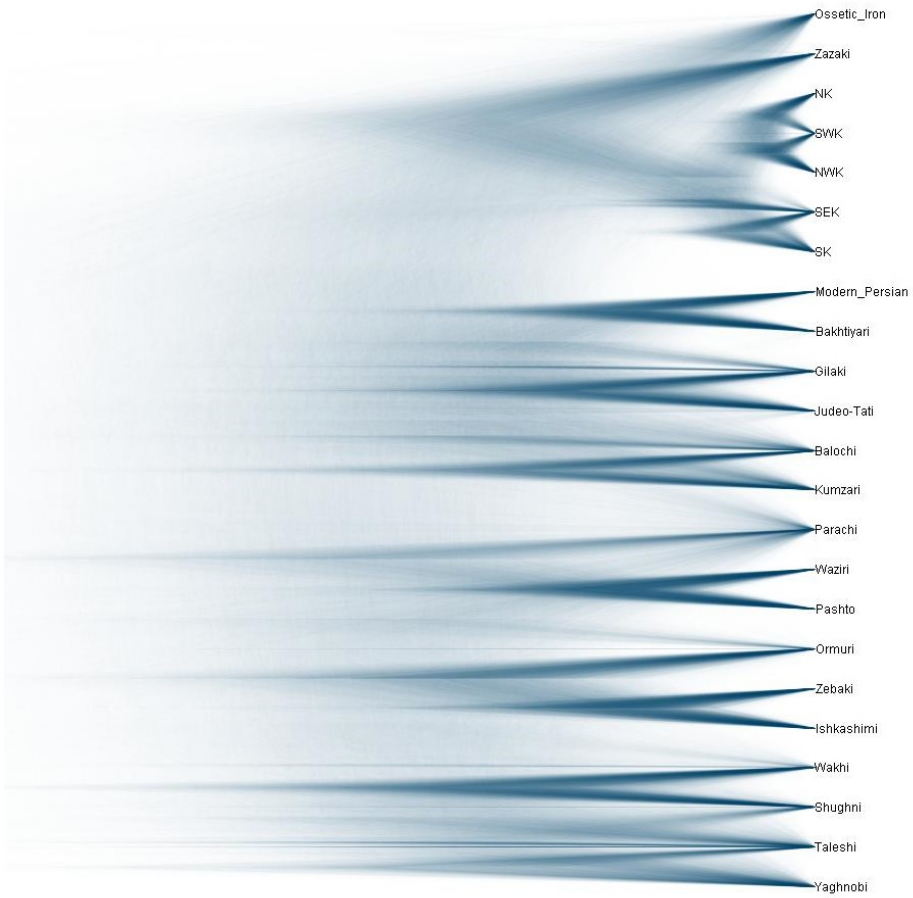
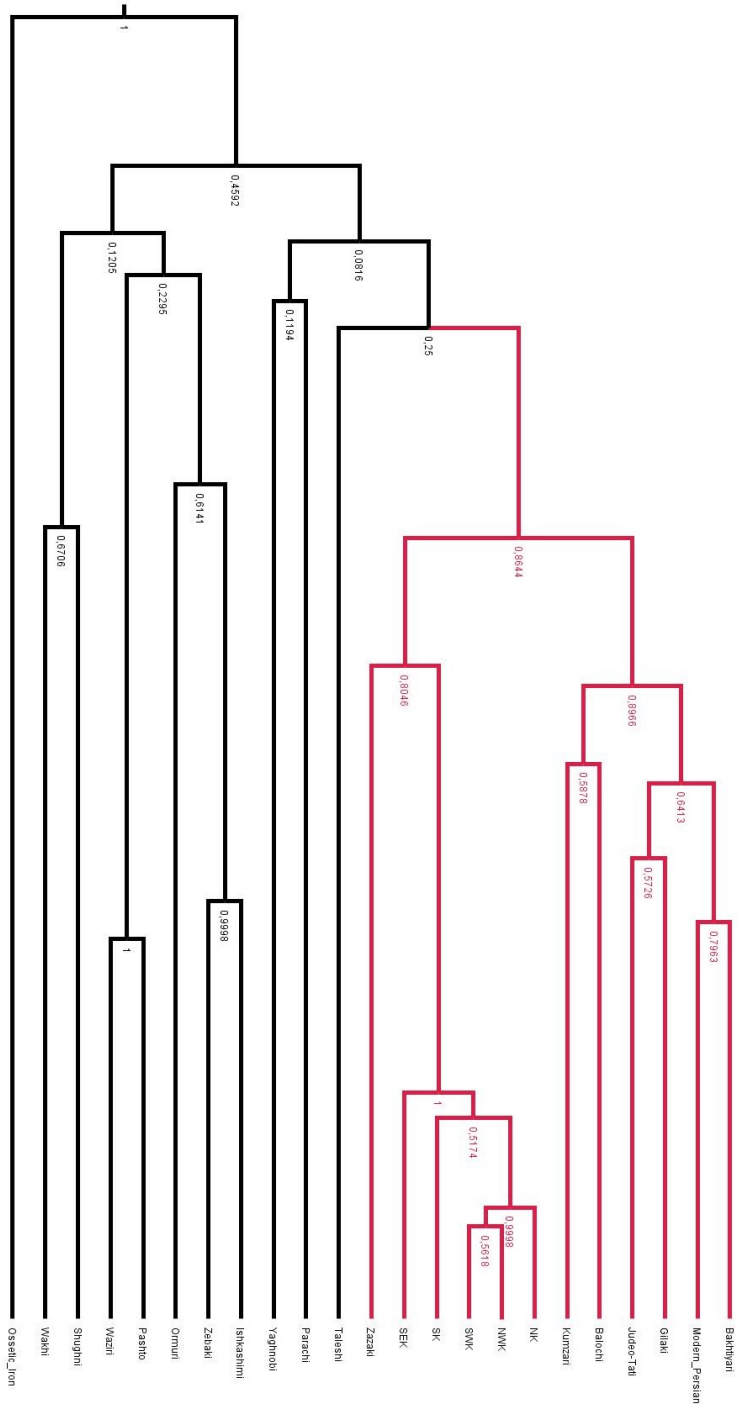

Ossetic_Iron
Zazaki
NK
SWK
NWK
SEK
SK
Modern_Persian
Bakhtiyari
Gilaki
Judeo-Tati
Balochi
Kumzari
Parachi
Waziri
Pashto
Ormuri
Zebaki
Ishkashimi
Wakhi
Shughni
Taleshi
Yaghnobi

Figure 4: Lexical only phylogeny as produced by covarion model



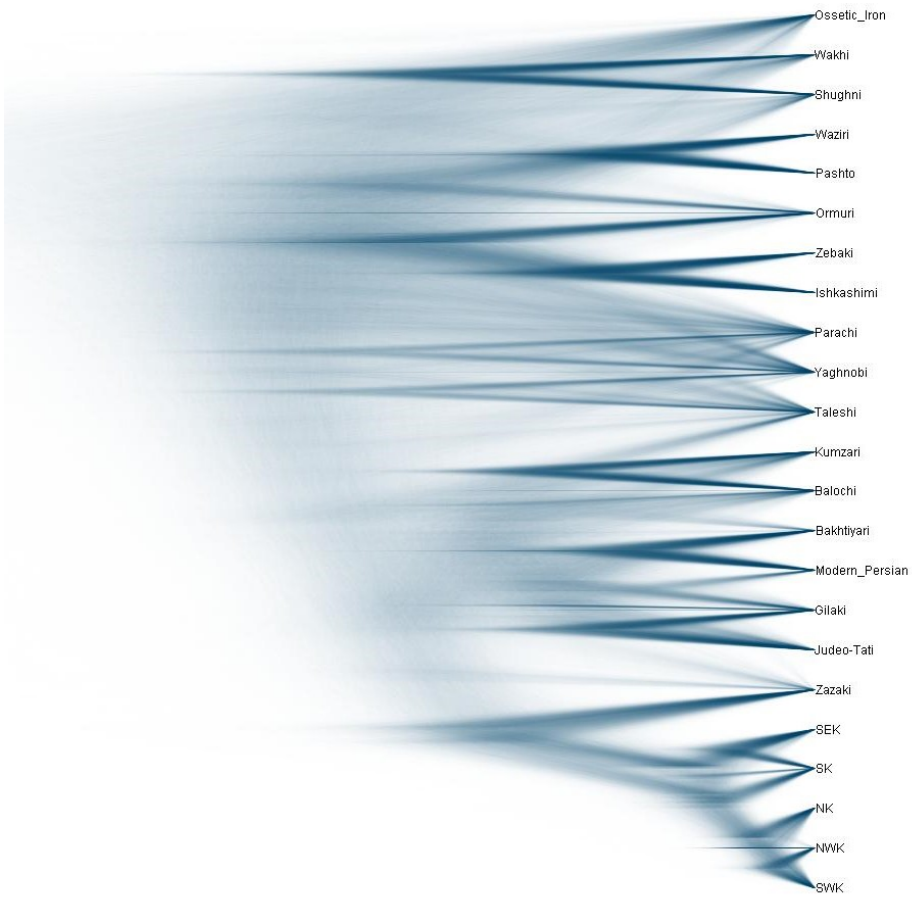
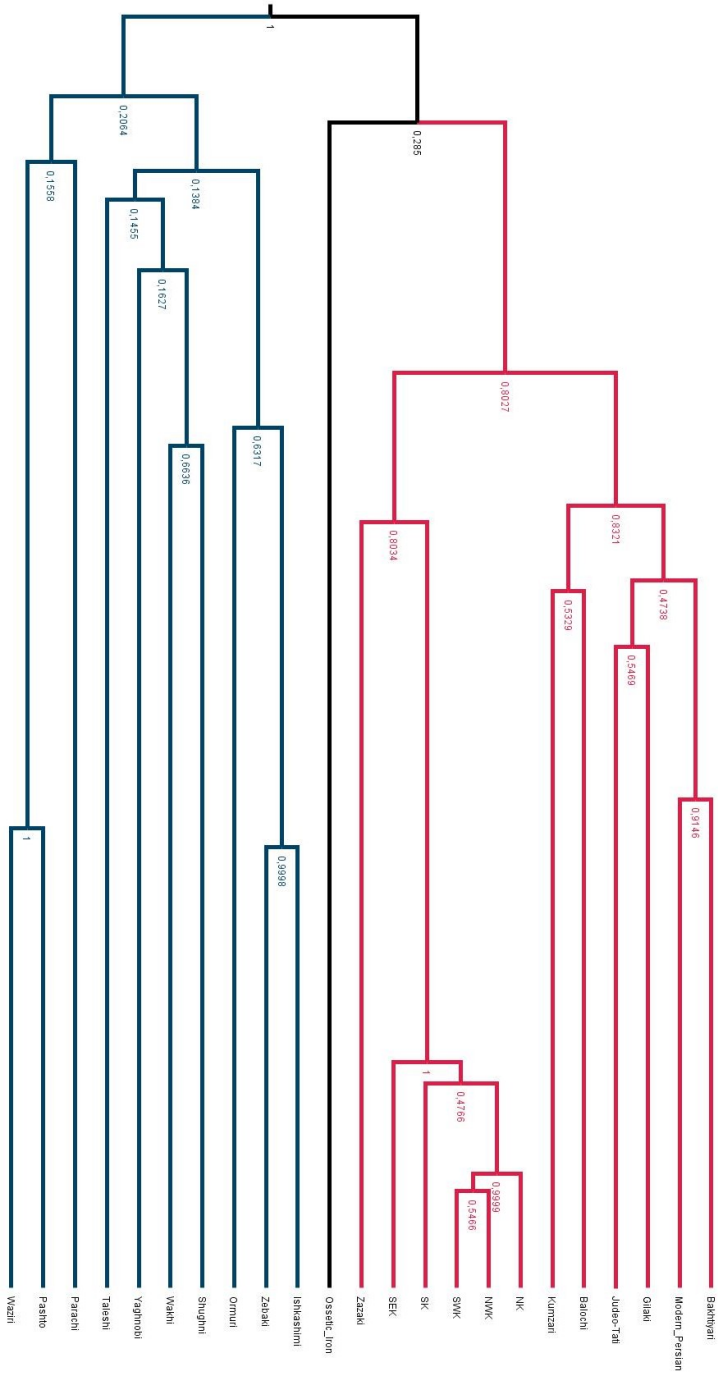
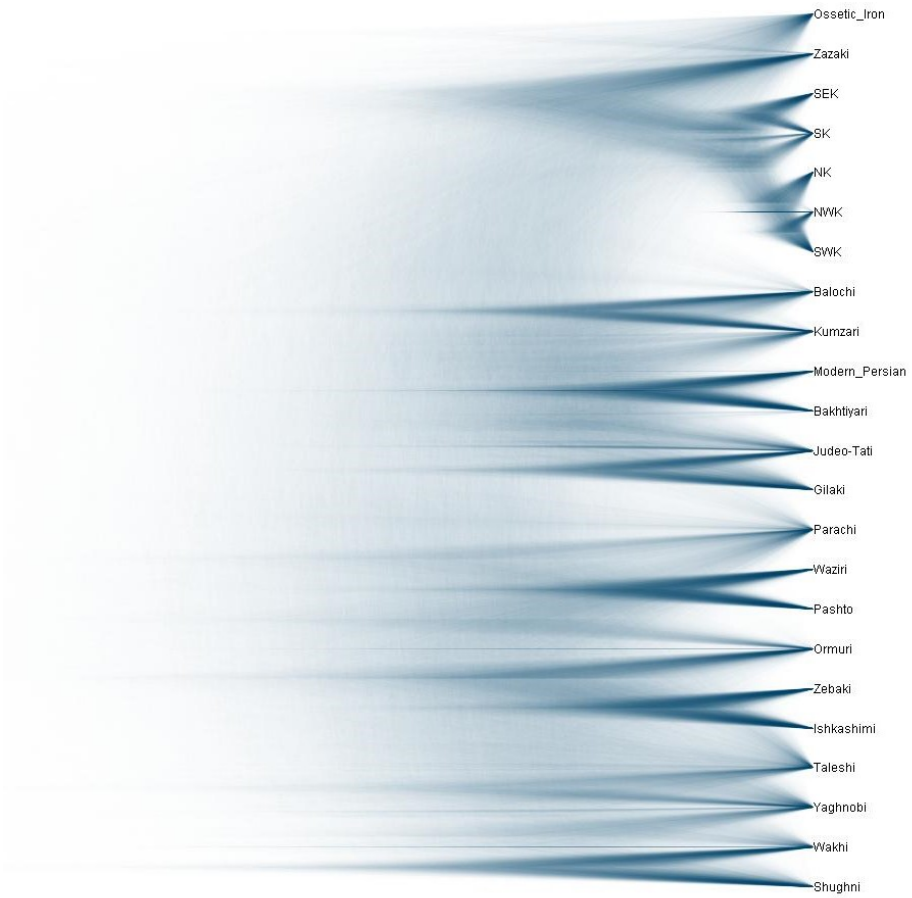

Ossetic_Iron
Wakhi
Shughni
Waziri
Pashto
Ormuri
Zebaki
Ishkashimi
Parachi
Yaghnobi
Taleshi
Kumzari
Balochi
Bakhtiyari
Modern_Persian
Gilaki
Judeo-Tati
Zazaki
SEK
SK
NK
NWK
SWK

Figure 5: Combined lexical and typological phylogeny as produced by covarion and BSVS models for lexical and typological data




Ossetic_Iron
Zazaki
SEK
SK
NK
NWK
SWK
Balochi
Kumzari
Modern_Persian
Bakhtiyari
Judeo-Tati
Gilaki
Parachi
Waziri
Pashto
Ormuri
Zebaki
Ishkashimi
Taleshi
Yaghnobi
Wakhi
Shughni