THE PROBLEM OF INTERPRETATION OF PHYLOGENETIC TREES

Valery Solovyev
Kazan Federal University
Timur Galeev
Kazan Federal University
Justus Liebig University of Giessen

Abstract. Phylogenetic algorithms have been used in a number of papers to describe the evolution of language families. In the paper the neighbor joining algorithm apply to the database of the Automated Similarity Judgment Program and results are compared with the common languages classification. A number of families have been considered in detail: North Caucasian languages, Turkic languages, Maya. In addition to recognized families, a hypothetical Nostratic macrofamily is also considered. When applying phylogenetic algorithms to databases, some errors occur. Possible causes of mistakes are analyzed, and a statement that mistakes are inevitable for phylogenetic algorithms is justified. The following main types of errors are identified. Languages in databases are represented as vectors of large dimension, while in the form of trees it is a one-dimensional structure. With decreasing dimension, the loss of information is mathematically unavoidable. Testing of one of the most popular phylogenetic algorithms – the algorithm of the neighbor joining – has been carried out, and it is shown that it gives an error in 13% of cases. Another source of error is the instability of phylogenetic algorithms – small (random) changes in the data can lead to a significant rearrangement of trees. A few recommendations on the methods of correct interpretation of results obtained via phylogenetic algorithms are proposed.

Keywords: Phylogenetic Algorithm, Evolution Trees, ASJP Database, North-Caucasian languages, Turkic Languages.

1. Introduction

An important trend in contemporary historical linguistics is the creation of databases of languages descriptions and subsequent applications of computer phylogenetic algorithms to build classifications of languages. However, the obtained results are not always satisfactory [1, 2]. Previous study [3] provides a tree of about one half of the world's languages constructed through the neighbor joining (NJ) algorithm from the database of the Automated Similarity Judgment Program (ASJP) data. Accordingly, lexical similarity registered in the tree is caused by four factors: (1) genetic or genealogical relationship of languages, (2) diffusion (language borrowing), (3) universal tendencies for lexical similarity such as onomatopoeia, and (4) random variation (chance). However, many cases of close location on the tree of languages and groups cannot be explained in such a way. In the present study we argue that there is another reason for the close location of languages on the tree – the mistakes of the algorithms for trees construction. Some ideas are proposed for reducing the influence of algorithm distortions. Different algorithms and different data (both ASJP-2 [3] and ASJP-3 [4]) are under consideration. ASJP-3 data are used unless otherwise indicated.

2. Methods

1.1. Examples of Phylogenetic Algorithms' Mistakes

Let us give a few examples of obvious NJ algorithm mistakes. The Greek language is situated on one branch together with the Nilo-Saharan language Koman and the Trans-New Guinea language Tanahmeran (Fig. 1 in Application). Borrowings are impossible in this case. Can it be by chance?

The Ukrainian language is situated not on the branch of East-Slavic languages, as it should be, but between a big branch, containing all other Slavic languages, and another, containing the Baltic languages (Fig. 2). The ASJP distance between Ukrainian and Latvian is 92.110 and the distance between Ukranian and Lithuanian is 90.940. The distances between Ukrainian and Slavic languages vary in the range from 50.990 to 68.640, with the smallest distance being to Belarusian: 50.990. Thus, the incorrect placement of Ukrainian is not a result of wrong ASJP-project data, but a consequence of an error produced by the NJ algorithm.

These are examples of wrong locations of separate languages, but whole groups are located incorrectly as well. Let us take the large group of Turkic languages. It is located far from Altaic languages and a branch which contains languages from Wakashan, Kuto, Nilo-Saharan and Yeniseian and looks very strange. There are many other examples.

1.2. Causes of Errors of Phylogenetic Algorithms: Non-stability

We suppose that a reason of close location of languages on the tree may be the mistakes of the NJ algorithm for tree construction. It is true that there is no evidence of the fact that NJ as well as any other phylogenic algorithm constructs trees in the right way. Let us mention a few reasons why one should be very careful with NJ-trees.

1). Generally, an evolutionary tree without mistakes is impossible. Languages in ASJP are characterized by 40 words, thus, using mathematical terms, they are represented as points in a 40-dimensional space of features. A tree can be represented as a symbol string in a bracket form, i.e. it is a 1-dimentional object. When projecting a 40-dimensional space onto a 1-dimensional one, distortions are inevitable.

2) Phylogenetic algorithms cannot only give wrong results but do it very often. For testing the NJ algorithm we generated 100 random trees (with 12 leaves) and constructed matrixes of distances for them. Then we applied the NJ

algorithm to these matrixes and compared the obtained trees with the initial ones. It turned out that NJ gave the right result only in 13% cases.

3). One cannot be sure that NJ is the best phylogenetic algorithm. It has been shown in the paper [2] that algorithm Unweighted Pair-Group Method Using Arithmetic Averages (UPGMA) gives better result than NJ for the Sumbanese languages. According to the Robinson-Foulds metric, the distance from the UPGMA tree to the consensus tree for the Sumbanese languages is 6, and the distance from NJ tree is 10.

4) Phylogenetic algorithms are notable for non-stability, i.e. a very small local change in the data can cause global reconstructions of the tree. Let us analyze a few examples, which illustrate point 4, comparing trees from ASJP versions 2 and 3.

I. North-Caucasian languages. Versions ASJP-2 [4] and ASJP-3 [3] have minimal difference, i.e. the two introduced dialects - Archi 2 and Bezhta 2.

As a result, the group of NorthWest Caucasian languages separated from other North-Caucasian languages in ASJP-3 and united with Tsimshianic, Botocudo, Maxakali, Pataxo and Yanomam. I suppose it is impossible to give linguistically relevant explanation of this facts. This is a display of pure non-stability of NJ algorithm.

One can receive some additional information about the NorthWest Caucasian languages from papers [5, 6].

II. Maya. A number of dialects are added to the new version: in particular Jacaltec Western appeared in Qanjobalan group except Jacaltec.

The Qanjobalan group moved to another part of Maya language tree in ASJP-3 tree (Fig. 3). In ASJP-2 [4] this group was situated close to the relative Chuj language, which is included to Greater Qanjobalan group according to well-known classification. In ASJP-3 tree Qanjobalan languages are situated together with Ixilan languages from the Eartern Mayan subgroup.

There is an article on Mayan by Cecil Brown on the project web-site, which explains lexical closeness of Mayan languages in ASJP-2 tree based on relationship and geographical proximity. Obviously it is time to give other explanations.

III. Turkic Languages. According to Ethnologue [6], the Turkic languages are divided into 5 subgroups: Eastern, Northern, Southern, Western and Bolgar, consisting of one language – Chuvash. No closer relationships among these groups is suggested in Ethnologue (Fig. 4).

The Khorasani language (with a lot of dialects) has been added to ASJP-3 as well as Turkish 2, Misher (Tatar dialect) и Crimean Tatar. It is significant that a complete reconstruction of the tree took place on a global level.

Separations from the root take place in ASJP-2 as follows: the Chuvash language at the beginning, the branch with the main part of Southern languages, the branch with the main part of Northern languages and the branch with complex mixture of the rest of languages (Fig. 5, left).

The Chuvash language also separates from the root in ASJP-3 at the beginning, but then the order is different: Turkish, Karaim, the branch with all Northern languages and the branch with all the other languages, Southern languages being located on the deepest level (Fig. 5, right).

It is interesting that there are similar problems and results while using the method of glottochronology with automatic construction of trees. In paper [8] there are two trees of Turkic languages (Fig. 6), constructed through the glottochronological method as specified by S. Starostin using the Starling algorithm. They differ in initial data: one of them uses 100-word lists with the majority of synonyms deleted.

Using unedited lists the sequence of branching is as follows: Chuvash, Northern, some Southern languages, and the rest. Using the edited lists the sequence is different: Chuvash, Southern, Northern (by two subgroups), and the rest. Thus, the sequence separations of the Southern and Northern branches differs due to small changes of data.

By means of the example of Turkic languages let us illustrate how one can reveal probable errors in the algorithm.

We shall analyze the location of the Khakas language in the ASJP-3 tree (Fig. 7).

Khakas is situated on one and the same branch with Sakha and Dolgan although according to papers by Starostin and other Russian scholars it belongs to a subgroup together with Altai and Shor. Can this be explained by borrowings? There are no borrowings from Sakha and Dolgan in the description of the Khakas language in the compendium “Languages of the World. Turkic languages”. Neither there are borrowings from Khakas in the languages. The geographical distance from Khakas to Sakha is more than 2400 km, while the distance between Khakas and Altai is only 420, and even smaller between Khakas and Shor (255).

If one consults the database of grammar features “Jazyki Mira – Languages of the World” [9], it turns out that the grammatical distance from Khakas to Sakha is 201 and 251 to Dolgan, while the distance from Khakas to Altai is 109 and 91 to Shor. Thus, one cannot find any features of probable proximity of the Khakas language with Sakha and Dolgan. Coming back to the data of ASJP-project we should notice that the lexical distance from Khakas to Sakha is 58.53 while it is smaller to Altai and Shor (43.12 and 55.72 respectively). Finally, taking the ASJP-2 tree, we can see that the Khakas is located in the same as Altai and Shor and far from Sakha and Dolgan (Fig. 8).

All this taken together brings us to the conclusion that there is an error of the NJ algorithm in the construction of the ASJP-3 tree for the Khakas language.

We would like to stress that the paper does not criticize the ASJP project, but only shows that one should analyze ASJP data directly rather than NJ trees. We need new algorithms for data analysis and/or new approaches to interpreting their results that reduce the influence of distortions for tree constructing.

3. Results and Discussion: Reducing the Influence of Algorithm Distortions

3.1. Reduction of the Amount of Languages. One possibility is to reduce the number of languages that are processed simultaneously.

North Caucasian Languages.

In complete ASJP trees (both -2 and -3) the Avar-Andic-Tsezic group is divided into Avar-Andic subgroup, which is united with Lak-Dargwa subgroup and Tsezic subgroup, being external to the other North-Eastern Caucasian languages [4]. When analyzing only North Caucasian languages, one can see (Fig. 9) that both subgroups are located together on one branch of the tree, corresponding to the genealogical classification [6] and their geographical proximity. Thus, the restriction of the set of languages by a certain area (local analysis) leads to a decreasing number of mistakes due to non-stability of the algorithm and more accurate trees are arrived at.

3.2. Consideration of Whole Groups, not of Separate Languages.

Nostratic Languages.

If one takes only Nostratic languages (Indo-European, Altaic, Uralic, Kartvelian, Dravidian, Afro-Asiatic) and builds the tree for those (Figure 10) the Greek language is found in its proper family. It may be reasonable to not construct a whole global World tree at once, but only trees for different parts of the world, dividing it into macroareas such as the Americas, Europe, North and Central Asia, North Africa and so on, or even into smaller areas.

Another possibility is take into consideration the time depth which provide the most adequate results. In [10] it was mentioned, that one should be very careful with the classification at great time depths. I consider such this view to be inertia of thinking. There is an opinion in classical historical linguistics that the reconstruction of relative connections is possible only at the depth of no more than 6,000 -10,000 years. However, if we take a set of Nostratic languages (whose age is about 15,000 years [11]) and construct a tree from ASJP data as in Figure 10 we get a result which corresponds quite well to the sort of tree that macro-comparativists following S. Starostin would construct.

Thus, there is no precise data which can prove that the quality of a tree reconstruction strictly depends on temporal depths. Other factors (including borrowings) are probably more significant and valid at any depth.

4. Summary

The results of phylogenetic algorithms applying depend upon the complexity of the evolutionary processes of each subgroup of languages. The situation with North-Caucasian languages is rather simple. Languages in the Northern Caucasus that are located geographically close to each other are generally also closely related. As a result the ASJP-tree correlates almost precisely with the established genealogical classification. At another extreme we find the Turkic languages, which have undergone more complex processes of evolution. There were numerable campaigns of conquest in their history that led to peoples moving and mixing. This naturally influenced the evolution of the languages and makes the determination of relationships more difficult.

Although the article was about NJ algorithm and ASJP base, analogous situation occurs for other phylogenetic algorithms and other large linguistic databases. This indicates the necessity of development of new approaches in using the phylogenetic algorithms in languages classification.

5. Conclusions

Our main conclusions are:

- phylogenetic algorithms are sensitive to small changes in initial data
- the NJ algorithm constructs trees with a large number of mistakes and one cannot be sure that it is the best algorithm
- one must learn how to analyze initial data of the ASJP-project using other methods
- local analysis usually gives more linguistically relevant results
- North-Caucasian languages are a simple case for analyzing ASJP-tree while Turkic languages is a complex one.

Thus, it is important not only to improve quality of databases, metrics of languages proximity and phylogenetic algorithms, but also to apply them properly and interpret their results correctly.

6. Acknowledgments

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University and state assignment of Ministry of Education and Science, grant agreement № 34.5517.2017/6.7

References

1. Suleri, J., and Cavagnaro, E. (2016). Promoting Pro-environmental Printing Behavior: The Role of ICT Barriers and Sustainable Values. Dutch Journal of Finance and Management, 1(1), 38. https://doi.org/10.20897/lectito.201638

2. Donwey S, Halmark B Cox M , Norquest P , Lansing S. Computational feature-sensitive reconstruction of language relationships: Developing the ALINE distance for comparative historical linguistic reconstruction. *Journal of Quantitative Linguistics*. 2008, 15(4), pp. 340-369.

3. Müller A., Wichmann S., Velupillai V., Brown C. H., Brown P., Sauppe S., Holman E. W., Bakker D., List M., Egorov O., Belyaev R., Mailhammer M., Urban H., Geyer A., Grant G.. ASJP World Language Tree of Lexical

Similarity, pp. 1-67. https://www.academia.edu/1935031/ASJP_World_Language_Tree_of_Lexical_Similarity_Version_3_July_2010_

4. Bernasconi, Andrés, and Emilio Rodríguez-Ponce. "Importancia de la gestión institucional en los procesos de acreditación universitaria en Chile." Opción 34.86 (2018): 20-48.

5. Danilova V., Makarova E., Polyakov V., Solovyev V. Frequency-Based Relevant Grammar Features of the Caucasian Languages. *Indian Journal of Science and Technology*, 2016, 9(11), pp. 1-10.

6. Galeev T. I., Solovyev V. D., Comparison of various quantitative measures of proximity of languages: North Caucasian languages. *Turkish online journal of design, art and communication*. - 2017. - Vol.7, - P.184-192.

7. Lewis M., Paul P., Gary F., Simons S., Charles D., Fennig F. Ethnologue: Languages of the World, Nineteenth edition. SIL International: Dallas, Texas, 2016.

8. Rance, N. E. (2009). Menopause and the human hypothalamus: evidence for the role of kisspeptin/neurokinin B neurons in the regulation of estrogen negative feedback. Peptides, 30(1), 111-122.

9. Polyakov V., Solovyev V. Komp'juternye modeli I metody v tipologii i komparativistike (Computational Models and Methods in Typology and Comparative Linguistics). Kazan: Kazanskiy Gosudarstvennyy Universitet. (in Russian). 2006, pp. 1-48.

10. Wichmann S., Holman E. W., Müller A, Velupillai V., List J. M., Belyaev O., Urban M., Bakker D. Glottochronology as a (non-)heuristic for genealogical language relationships. *Journal of Quantitative Linguistics*. 2010, 17(4), pp. 303-16.

11. The Tower of Babel. An International Etymological Database Project. http://starling.rinet.ru/babel.php?lan=en. Access: 15.05.2018

APPLICATIONS

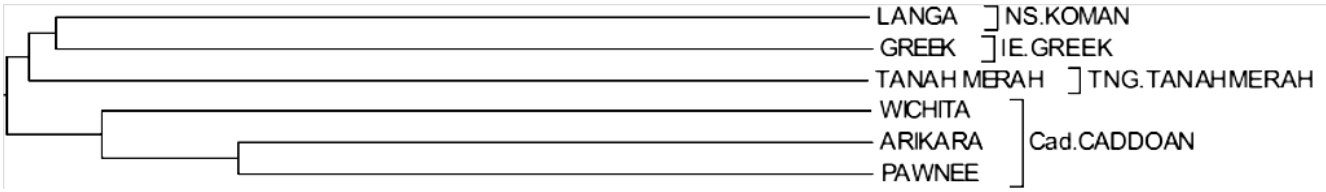

Figure 1. Fragment of tree including Greek

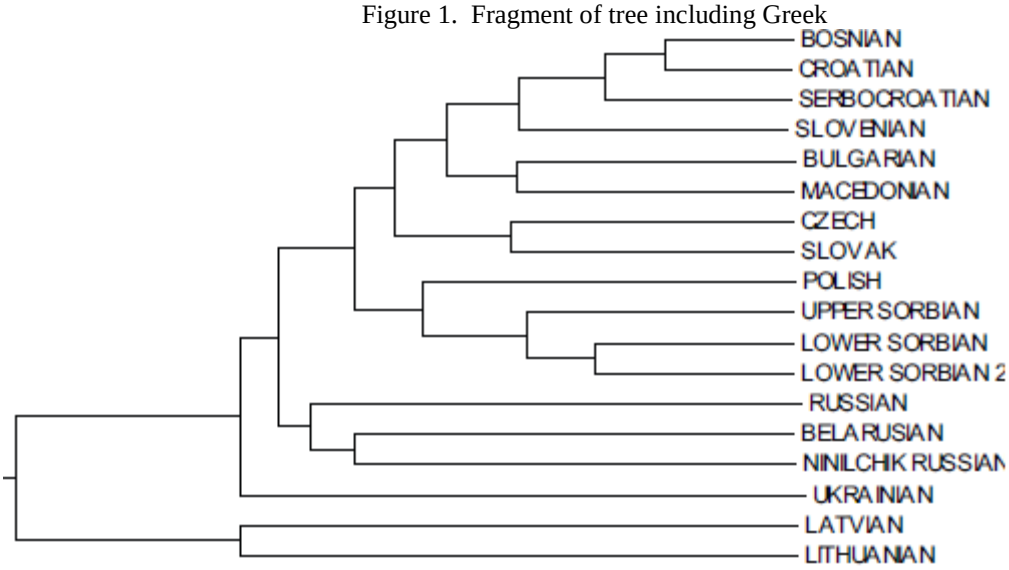

Figure 2. Balto-Slavic branch

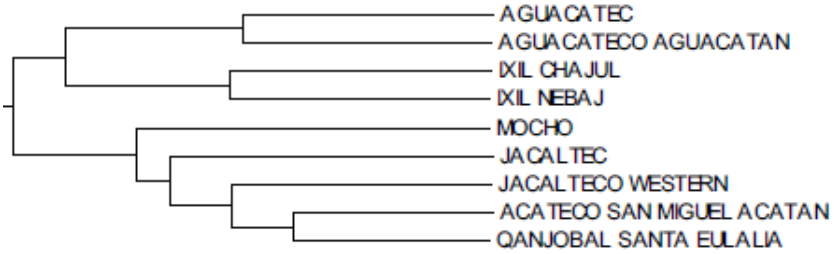

Figure 3. ASJP-3 Mayan tree

Figure 4. Turkic Languages according to Ethnologue

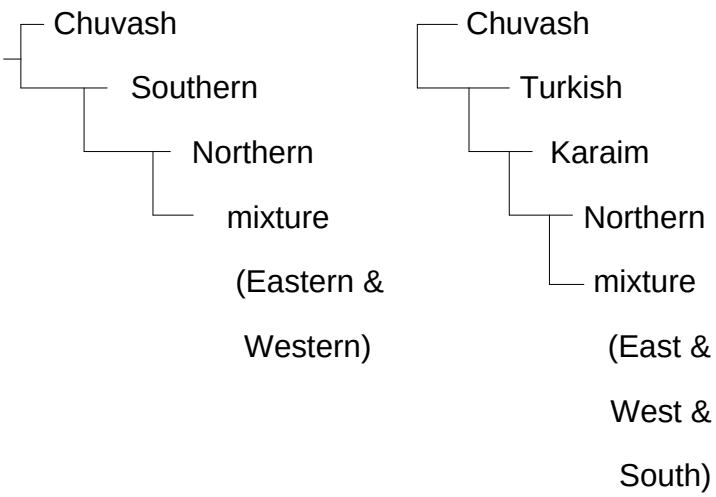

Figure 5. Main branches of Turkic tree in ASJP-2 (left) and in ASJP-3 (right)

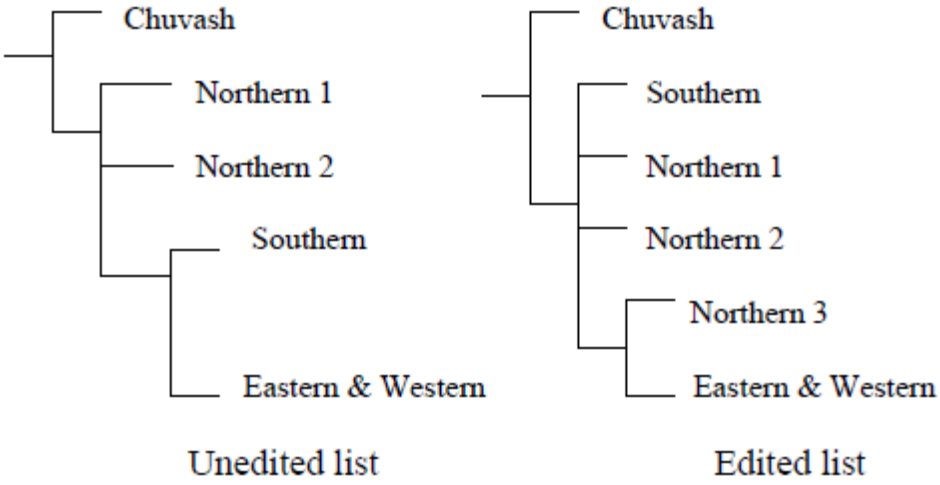

Figure. 6. Main Turkic branches (Glottochronology Method)

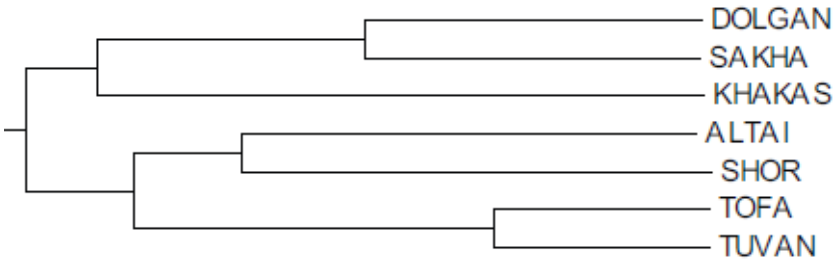

Figure. 7. Fragment of ASJP-3 with Khakas

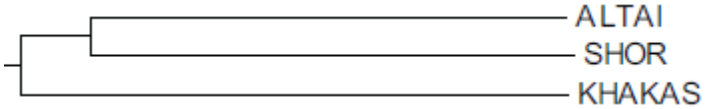

Figure. 8. Fragment of ASJP-2 with Khakas

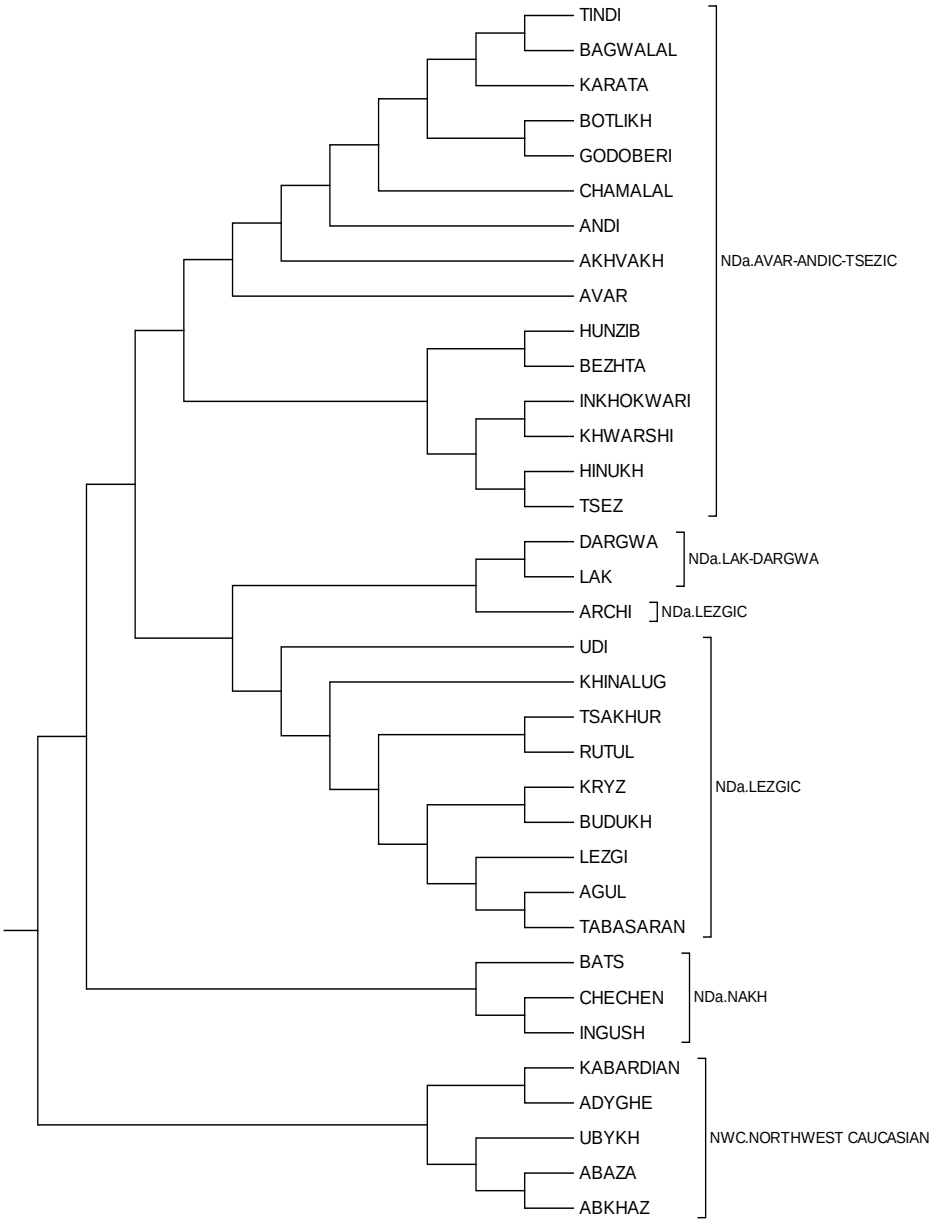

Figure. 9 North Caucasian languages only

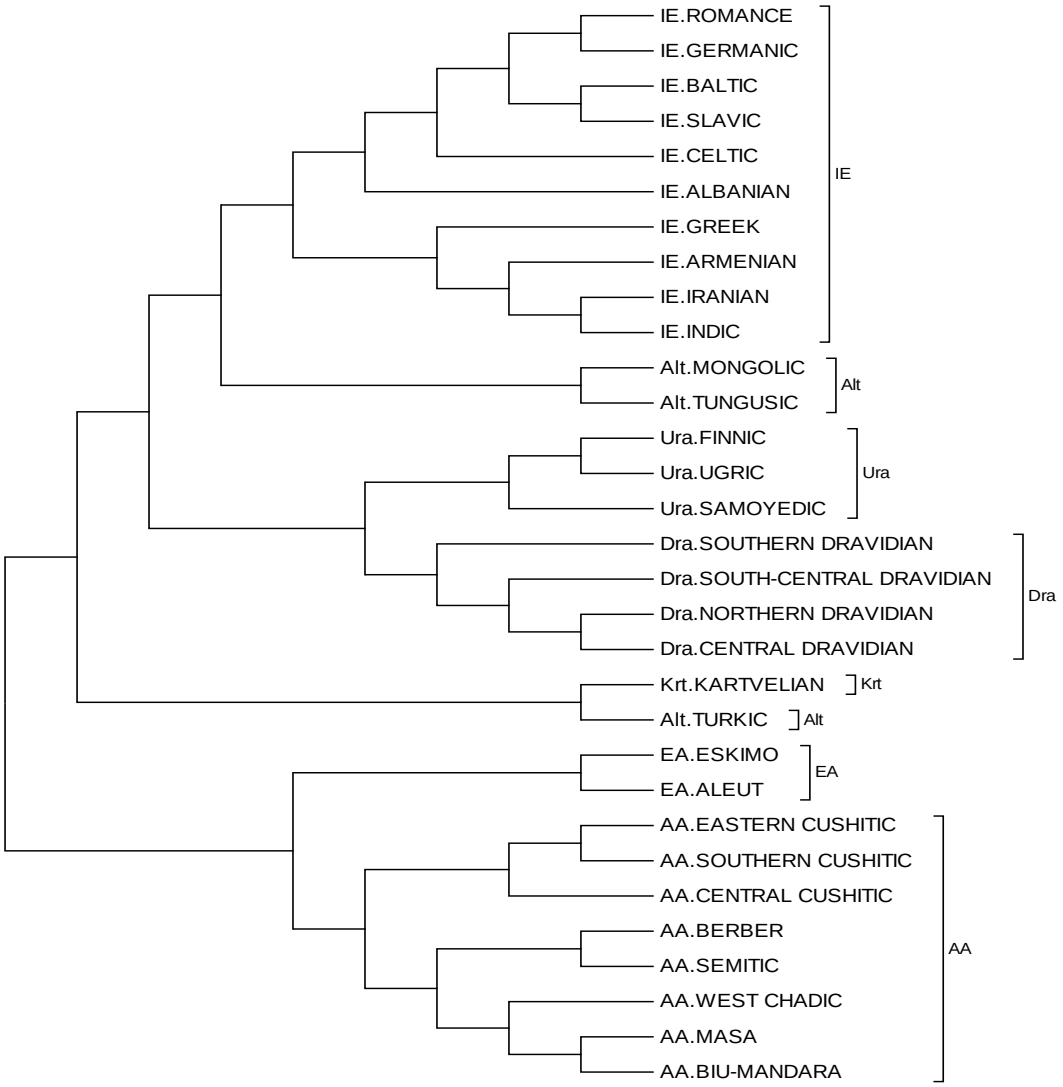

Figure. 10. Nostratic Languages