

# The Problem of Interpretation of Phylogenetic ASJP-trees

Valery Solovyev

Russia, Kazan University, maki.solovyev@mail.ru

In [1] there has been shown a tree, constructed by NJ algorithm on ASJP data. It mainly reflects the classification of languages based on comparative-historical method. According to the authors, 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 can not be explained in such a way. For example, the Greek language is situated on the tree close to Mascoian languages, which are the languages of Paraguay Indians. Borrowings are impossible in this case and onomatopoeia is not traced. In paper [1] the attention is focused on the fact that the influences of chance can be essential for smaller groups. Let us take rather large group of Turkic languages. It is located on the same branch with the Lower Sepik languages of New Guinea that looks very strange. There are many other examples. Such number of coincidences is statistically improbable.

I suppose that except the four ones mentioned above there is another reason of close location of languages on the tree, i.e. the mistakes of the algorithm NJ for trees construction. It is true that there is no evidence of the fact that NJ as well as any other phylogenetic algorithm constructs trees in the right way.

It can be obvious from the common sense, that the situation without mistakes is impossible. Languages in ASJP are characterized by 40 words, thus, using mathematical terms, they are represented as dots in 40-dimensional space of features. A tree can be represented as a string in a bracket form of writing, i.e. it is 1-dimensional object. While projecting 40-dimensional space on 1-dimensional one, distortions are inevitable.

It raises the problem of reducing the influence of distortions for tree constructing. One of the probable ideas consists in reducing the number of simultaneously-processed languages. So if one takes only Nostratic languages (Indo-European, Altaic, Uralic, Kartvelian, Dravidian, Afrasian) the Greek language and the Turkic group are found in their families. It may be reasonable not to construct a whole global World tree at once, but by parts, having divided the World into macroareas, such as America, Europe, North and Central Asia, North Africa and so on, or even into smaller ones.

Another problem is a temporal depth, which provides adequate results. In [2] it was mentioned, that one should be very careful with the classification on the great temporal depth. I consider such views to be inertia of thinking. There is an opinion in classical historical linguistics that reconstruction of relative connections is possible on the depth of no more than 6,000 -10,000 years. However, if we take a set of Nostratic languages (age of 15,000 years [3]), the NJ algorithm constructs a tree on ASJP data, which perfectly corresponds to the constructed tree in macro-comparativistics (Starostin's school, [3]).

At the same time a tree for Turkic languages of 2,000 years depth abounds in incorrect classifications. So there are representatives of three sub-groups on one branch: Kypchak languages (Kumyk, Karachay-Balkar, Nogai), a Karluk language (Uzbek) and a Khalaj language (Khalaj). Oguz language (Salar) on another branch is united with Siberian languages - Altai, Shor, Khakas. Yakut languages (Dolgan, Sakha) and Siberian languages (Tofa, Tuvan) are located together on the third branch. Kypchak languages are distributed in three places on the tree. Oguz languages are found in two places and Siberian languages in two.

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

The work has been completed with support of Russian Ministry for education and science, grant No 10-06-00087-a.

## References

1. Müller, André, Viveka Velupillai, Søren Wichmann, Cecil H. Brown, Pamela Brown, Eric W. Holman, Dik Bakker, Oleg Belyaev, Dmitri Egorov, Robert Mailhammer, Anthony Grant, and Kofi Yakpo. 2009. ASJP World Language Tree of Lexical Similarity: Version 2 (April 2009).
2. Wichmann, Søren, Eric W. Holman, André Müller, Viveka Velupillai, Johann-Mattis List, Oleg Belyaev, Matthias Urban, and Dik Bakker. Glottochronology as a (non-)heuristic for genealogical language relationships. <http://email.eva.mpg.de/~wichmann/GlottoHeurUpload.pdf>
3. The Tower of Babel An International Etymological Database Project. <http://starling.rinet.ru/babel.php?lan=en>. 2010.