ASJP World Language Tree of Lexical Similarity: Version 5 (September 2021)

by

Müller, André, Viveka Velupillai, Søren Wichmann, Cecil H. Brown, Eric W. Holman, Sebastian Sauppe, Pamela Brown, Harald Hammarström, Oleg Belyev, Johann-Mattis List, Dik Bakker, Dmitri Egorov, Matthias Urban, Robert Mailhammer, Matthew S. Dryer, Evgenia Korovina, David Beck, Helen Geyer, Pattie Epps, Anthony Grant, and Pilar Valenzuela.

The World Language Tree graphically illustrates relative degrees of lexical similarity holding among 7996 of the world's languages and dialects (henceforth 'doculects', when referring to the specific variants of the database and 'languages' when speaking in general terms) currently found in the ASJP database. ASJP stands for Automated Similarity Judgment Program. Languages branched more closely together on the ASJP tree are lexically more similar than those branched less closely together. While most lexical resemblance charted in the tree almost certainly is related to genetic affiliation, closely branched languages cannot routinely be assumed to be closely genetically associated since lexical resemblance can be due to factors other than genetic relatedness (see below).

The tree is generated through use of the Neighbour-Joining computer algorithm originally designed to depict phylogenetic relationships in biology (Saitou & Nei 1987). This is implemented in MEGA 7 (Kumar et al. 2016),[1] the software that we use. The algorithm is applied to a matrix of lexical similarity scores based on Levenshtein (or edit) distances holding between all possible pairs of the ~8000 doculects (for details about this, including how we modify the Levenshtein distances for our purposes, see Bakker et al. 2009: 169). All doculects of the database are compared to one another with respect to lexical similarity relating to their words for 40 referents determined statistically in Holman et al. (2008) to be most stable among core vocabulary used in the tradition of lexicostatistical analysis. The tree is unrooted, but organized around a midpoint, i.e., the point which is equidistant between the two most lexically dissimilar doculects in the network. The doculect names used are normally simply those of the sources consulted. The sources, as well as other metadata is provided at the ASJP website.[2]

---

[1] http://www.megasoftware.net/

[2] https://asjp.clld.org/

Four factors influence lexical similarity registered in the tree: (1) genetic or genealogical relationship of languages, (2) diffusion (borrowing), (3) universal tendencies for lexical similarity such as onomatopoeia, and (4) random variation (chance).

Languages branched closely together on the tree may be so because of strong lexical similarity produced by any one or a combination of the four factors. Genetic relationship would appear to be the most dominant factor accounting for close branching, followed next by diffusion. Universal tendencies and chance are less significant contributors to close branching than either genetic relationship or diffusion, but nonetheless clearly contribute to the overall structure of the tree. The effect of diffusion is lessened somewhat since known loanwords are excluded from the similarity calculations, but these known loanwords were not identified through extensive research and would only represent a very small fraction of the actual loanwords.

Typically, all languages of non-controversial language families such as Austro-Asiatic, Uralic, or Mayan, are respectively branched together on the tree. When some languages of a non-controversial family are not found branched together, this is because they are substantially lexically different from other members of their family despite unambiguously belonging to that family. Occasionally, a language can be so lexically different from co-members of its family that it is found branched more closely with some language or languages with which it is not genetically related at all, usually because of chance lexical similarity or similarity due to borrowing. (When such languages are geographically remote from one another, chance usually explains close branching.)

Typically, branching accords closely with genetic subgroups recognized by experts within non-controversial language families. When branching is not isomorphic with genealogical subgrouping, this often reflects diffusion among languages of the family promoted by language contact. Thus, when used in conjunction with expert classifications of non-controversial language families, the tree can be helpful in calling attention to historical relationships (contact) among genetically related languages that sometimes might not be otherwise apparent.

The tree may also suggest relationships heretofore not noticed among languages that may be profitably investigated. For example, if two languages not known to be related in any way are found together on a terminal branch, this may indicate a relationship between them entailing either inheritance or contact, especially if they are not geographically remote from one another. If the two languages are geographically distant, their close lexical similarity is more likely explained by chance than by either inheritance or diffusion. Also, language

isolates may join one another on a terminal branch because they have nowhere else to go in the tree, creating the illusion that exciting, new far-flung relations may be in evidence. One should be cautious in the interpretation of these cases.

For technical reasons relating to software limitations this version of the ASJP World Language Tree really consists of two trees: one for Australian, 'Papuan', and Austronesian languages, and one for the rest. We exclude languages documented only before 1700 A.D., as well as doculects for which only less than 28 items on the 40-item list were available.

References

Bakker, Dik, André Müller, Viveka Velupillai, Søren Wichmann, Cecil H. Brown, Pamela Brown, Dmitry Egorov, Robert Mailhammer, Anthony Grant, and Eric W. Holman. 2009. Adding typology to lexicostatistics: a combined approach to language classification. *Linguistic Typology* 13: 167-179.

Dryer, Matthew S. & Haspelmath, Martin (eds.). 2011. *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library. Available online at http://wals.info/ Accessed on 2013-10-22.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker. 2008. Explorations in automated lexicostatistics. *Folia Linguistica* 42.2: 331-354.

Kumar Sudir, Glen Stecher, and Koichiro Tamura. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* 33: 1870-1874.

Saitou, Naruya and Masatoshi Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4: 406-425.

Tamura, Koichiro, Glen Stecher, and Sudir Kumar. 2021. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution* 38(7): 3022-3027.