《International Symposium》
Let's Talk about Trees

National Museum of Ethnology, Osaka, Japan. February 10, 2013

Abstract

[Various Tree models in Linguistics]
Who's to blame for the black sheep?
A study of reticulation and classification performance of distance-based language trees at the level of individual taxa

Søren WICHMANN

Max Plank Institute for Evolutionary Anthropology

The Automated Similarity Judgment Program (ASJP) uses a version of the edit distance (a.k.a. Levenshtein distance) to calculate distances among languages using 40-item word lists. Because the database covers well over one half of the world's languages it is possibly to generate distance matrices for languages from almost all the world's language families and study their properties. Moreover, the matrices can be compared with trees in the literature, such as those of Ethnologue. This paper looks at two properties or distance matrices. One is reticulation, i.e. conflicting phylogenetic signals, and the other is the goodness-of-fit between ASJP distance matrices and standard classifications. These properties have been studied before at the level of whole families, but here I show how they can also be studied at the level of individual languages. Using the example of Austronesian I show how such a look at the behavior of individual languages can be helpful. Previous studies have shown that the fit between ASJP distances and expert classifications varies a lot from family to family, but the source of this variation in fit is not well understood. Correlation tests have shown that there is a tendency for poorly fitting families to also be more reticulate, indicating that when ASJP results differ from the experts' classification the method is also experiencing problems in fitting the data to a tree model. But in many case, including several Austronesian subgroups, ASJP distances are not very reticulate, although the fit with expert classifications are poor. It is suggested that in such cases it may be worthwhile looking into whether the expert classification are actually well supported or whether ASJP in these cases could be more adequate.