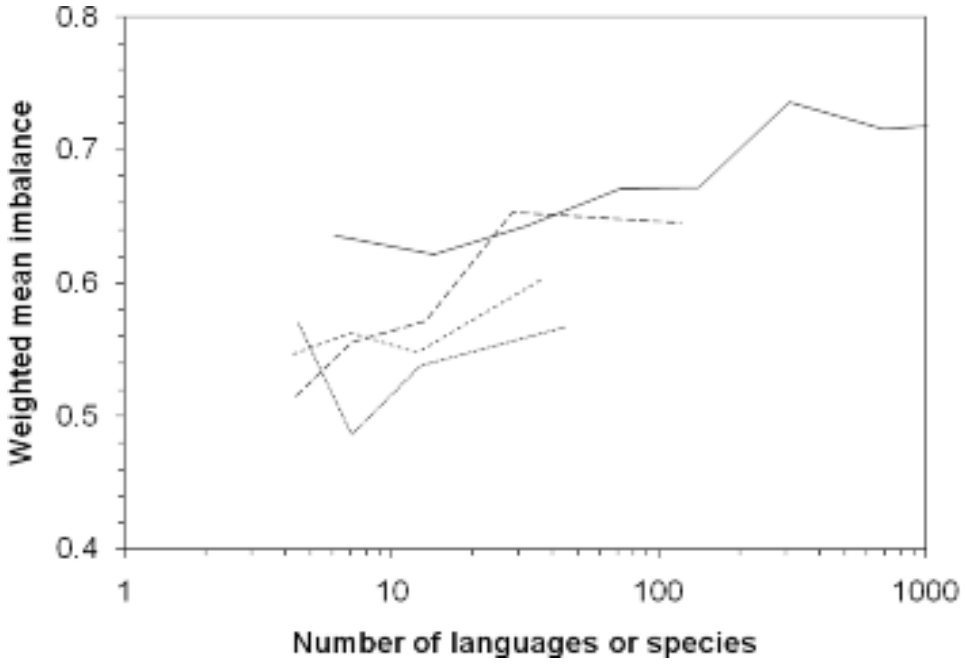
Eric W. Holman

Do languages originate and become extinct at constant rates?

Morris Swadesh explored the possibility that the words for a given meaning in basic vocabulary are replaced at a stochastically constant rate over time. Other possibilities for constancy are the rate at which languages split to form additional daughter languages, and the rate at which languages become extinct without descendants. The constancy of these rates defines a simple birth and death process, which has already been applied to the origination and extinction of species and used to predict the shape of phylogenetic trees in biology. One standard measure of tree shape is weighted mean imbalance as defined by Purvis et al., 2002, *Journal of Theoretical Biology* 214:99-103. For any binary node in a tree, imbalance takes the value 0 if the two branches are as balanced as possible (with equal numbers of languages or with numbers differing by only one), and it takes the value 1 if the branches are as unbalanced as possible (with one language on one branch and all the rest on the other), and it takes intermediate values for intermediate degrees of balance. According to the birth and death model, weighted mean imbalance has expected value .5, independent of the birth and death rates and the total number of languages on the two branches.



The chart shows weighted mean imbalance as a function of the total number of languages (or species) on the branches, for four sets of phylogenetic trees. The dotted line refers to language trees in the Ethnologue classification, which were constructed manually. The long-dashed line refers to trees constructed by the neighbor-joining algorithm from language similarities produced by the Automatic Similarity Judgment Program (ASJP, see http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm) based on Levenshtein distances in a 40-item Swadesh lexicostatistical list. The short-dashed line also refers to the Ethnologue trees but includes only the languages that are also attested in ASJP. For comparison, the solid line refers to a set of published biological phylogenetic trees, which were constructed automatically from morphological or genetic data. Each curve is significantly above .5 on average, contradicting the birth and death model. The ASJP and biological curves, which have the most binary nodes, also have significantly positive slopes.

The usual explanation for unbalanced biological trees is that species on some branches are better adapted to their environment and therefore have higher origination rates or lower extinction rates. Alternative explanations are explored for languages, which are not thought to differ in adaptation. Average number of speakers per language is no greater on branches with many languages than on branches with few. There is, however, indirect evidence for temporary reductions in rates of origination and extinction associated with the adoption of writing.