Automated dating of the world's language families based on lexical similarity

Holman E., Brown C., Wichmann S., Müller A., Velupillai V., Hammarström H., Sauppe S., Jung H., Bakker D., Brown P., Belyaev O., Urban M., Mailhammer R., List J., Egorov D.
Kazan Federal University, 420008, Kremlevskaya 18, Kazan, Russia

Abstract

This paper describes a computerized alternative to glottochronology for estimating elapsed time since parent languages diverged into daughter languages. The method, developed by the Automated Similarity Judgment Program (ASJP) consortium, is different from glottochronology in four major respects: (1) it is automated and thus is more objective, (2) it applies a uniform analytical approach to a single database of worldwide languages, (3) it is based on lexical similarity as determined from Levenshtein (edit) distances rather than on cognate percentages, and (4) it provides a formula for date calculation that mathematically recognizes the lexical heterogeneity of individual languages, including parent languages just before their breakup into daughter languages. Automated judgments of lexical similarity for groups of related languages are calibrated with historical, epigraphic, and archaeological divergence dates for 52 language groups. The discrepancies between estimated and calibration dates are found to be on average 29% as large as the estimated dates themselves, a figure that does not differ significantly among language families. As a resource for further research that may require dates of known level of accuracy, we offer a list of ASJP time depths for nearly all the world's recognized language families and for many subfamilies. © 2011 by The Wenner-Gren Foundation for Anthropological Research. All rights reserved.