

**George Starostin**

Russian State University for the Humanities

### **Automated vs. manual lexicostatistics: the mutual benefits**

As a representative of the Moscow school of comparative linguistics, there are three things about the ASJP Project that I find extremely useful and worthy of expanding and publicizing. The first is its acknowledging of *lexicostatistics* as a powerful — if not necessarily definitive — tool for language classification; the second — its firm resolve to arrive, if only on a preliminary level, at a global picture of the world's languages that could work as a general framework for further individual research; the third is, of course, its reliance on formalized automated algorithms in order to avoid subjectivity, a common plague of historical linguistics, especially in those linguistic areas where little proper historical research has been carried out so far.

My main commentary for this talk will dwell on the third point — the degree of validity of automatic lexicostatistical algorithms for their stated purposes; this is of particular interest to me, since parallel implications of automatic and manual lexicostatistical analysis are currently at the core of the «Evolution of Human Language» project (Santa Fe Institute) in which I am an active participant. Based on the comparison of (a) some results displayed on the ASJP tree of lexical similarity, (b) results of automatic calculations performed on EHL's preliminary lexicostatistical tree of the languages of Eurasia, (c) results of manual calculations performed on the same tree, and illustrating all this with actual linguistic examples, I will try to make the following points:

(1) Automated classificatory methods work best when certifying the integrity of «mid-size linguistic taxa» (MSLT) that are generally not older than  $\approx 6,000$  years and do not have any additional «relatives» on a higher, but chronologically not distant, level;

(2) Quality of results depends significantly on the size of the MSLT in question; language isolates with distant nearest relatives run a very high risk of being misclassified;

(3) Unequivocally relying on the results of *internal* classification of the languages that constitute the MSLT is not recommendable, since they may be seriously skewed by relying too much on phonetic rather than lexical innovation;

(4) Credible hypotheses on macro-relationship between several MSLTs, at this point in time, cannot be arrived at by relying entirely on available automated algorithms, due to their inability to distinguish relevant data from «noise» on deep time levels;

(5) They can, however, be advanced through manual lexicostatistical techniques, relying on strict procedures that involve reconstruction of intermediate proto-wordlists with step-by-step elimination of non-diagnostic lexical material — such as are currently developed within EHL. It is also possible that in the future we will be able to arrive at simulating such procedures on an automatic level as well.