Filippo Petroni and Maurizio Serva

Indo-European and Austronesian trees reconstruction

Languages evolve in time according to a process in which reproduction, mutation and extinction are all possible. This is very similar to haploid evolution for asexual organisms or for mtDNA of complex ones. Exploiting this similarity, it is possible, in principle, to verify hypothesis concerning the relationship among languages and to reconstruct their family tree. The key point is the definition of the distance among pairs of languages in analogy with the genetic distance among pairs of organisms. Distances can be evaluated comparing grammar and/or vocabulary but while it is difficult, if not impossible, to quantify grammar distance, it is possible to measure a distance from vocabulary differences. The method used by glottochronology, computes distances from the percentage of shared ``cognates'' which are words with a common historical origin. The weak point of this method is that subjective judgment plays a relevant role. Here we define the distance of two languages by considering a renormalized edit distance among words with same meaning and averaging on the two hundred words contained in a Swadesh list. In our approach the vocabulary of a language is the analogous of DNA for organisms. The advantage is that we avoid subjectivity and, furthermore, reproducibility of results is granted. We apply our method to the Indo-European and the Austronesian group considering, in both cases, fifty different languages. The two trees obtained are, for many aspects, similar to those found by glottochronologists with some important differences concerning the position of few languages. In order to support these different results we separately analyze the structure of distances of these languages with respect to all the others.