A quantitative study on Chinese dialects

Feng Wang, James Minett & Christophe Coupé

Three algorithms are used to classify Chinese dialects, two of them are distance based, and the third is character based. Our aim is to test which algorithm will be the optimal one for genetic classification. Moreover, the best topology of Chinese dialects will be suggested.

The two distance-based methods calculate the differences of shared cognates and some co-occurrences of phonological features in Chinese dialects, respectively. The first method assumes that the number of shared cognates in basic words will represent the genetic distance, but which basic word list is suitable is controversial. The second method explore whether rare co-occurrences or frequent co-occurrences of phonological features will represent the genetic relationship between languages. The third method will be based on the information of shared innovations.

The three methods will be tested in Chinese dialects. Some well-defined taxa, for example, Jinanese and Pekineses as Mandarin, Shanghainese and Suzhounese as Wu dialect, will be taken as the criterion to test the three algorithms. The algorithm failed to establish these taxa will be considered unsuitable for genetic classification for Chinese dialects. Another indicator for optimality of classifications is the stability of the topologies derived from the algorithms. In this test, the standard dialects for Seven Major Chinese dialects will be the fixed items, several other dialects whose position is quite sure are taken as optional items. With the change of optional items, the algorithms will produce different topologies for Seven Major Chinese dialects. The differences between topologies produced by each algorithm will be calculated; the smaller one will indicate the optimality of the algorithm. Namely, a program in progress will allow us to generate pseudo-random data for both lexicostatistical and character-based classification methods for an arbitrary number of languages and characters under simple conditions of borrowing. The pseudo-random data with known topologies will be used to test the performances of each method under carefully controlled conditions.

The most interesting result of our preliminary study is that the topologies produced by the algorithms strongly suggest the first division in Chinese will be Southern Chinese (Min and Hakka) and Northern Chinese (Pekinese, Xiang, Gan and Wu), which will attract some new thoughts about the formation of Chinese dialects.