Reconstructing the Evolutionary History of Chinese Dialects

Esra Erdem Institute of Information Systems, Vienna University of Technology, Vienna, Austria
esra@kr.tuwien.ac.at
Feng Wang Department of Chinese Language and Literature, Peking University, Beijing, China
wfwf@pku.edu.cn

Evolutionary relations between languages based on their shared characteristics can be represented as a phylogeny --- a tree where the leaves represent the extant languages, the internal vertices represent the ancestral languages, and the edges represent the genetic relations between the languages. On the other hand, languages not only inherit characteristics from their ancestors but also sometimes borrow them from other languages. Such borrowings can be represented by additional non-tree edges, turning a phylogeny into a phylogenetic network. With this motivation, we reconstruct the evolutionary history of languages in two steps: first we compute a plausible phylogeny with a minimal number of incompatible characters, and then we turn this phylogeny into a perfect phylogenetic network, by adding a small number of lateral edges, so that all characters are compatible with the network. For both steps, to formulate the problems and to solve them, we use answer set programming --- a new form of declarative programming. This method has been successfully applied to reconstruct the evolutionary history of Indo-European languages. In the following we summarize its application to reconstruct the evolutionary history of Old Chinese and the following 23 Chinese dialects: Guangzhou, Liancheng, Meixian, Taiwan, Xiamen, Zhangping, Fuzhou, Nanchang, Anyi, Shuangfeng, Changsha, Beijing, Yuci, Taiyuan, Ningxia, Chengdu, Yingshan, Wuhan, Ningbo, Suzhou, Shangai 1, Shangai 2, Wenzhou.

We have started with a dataset consisting of 200 lexical characters (the Swadesh wordlist), each with 1--24 states. We have obtained from this dataset, the part that is relevant to our computations: 148 informative characters, each with 2--6 essential states. On the other hand, we have identified some domain-specific information that, when added to our program, prevents the generation of some implausible phylogenies and implausible phylogenetic networks. For instance, there is a well-known classification of the dialects above into the following 7 groups: Yue, Hakka, Min, Gan, Xiang, Mandarin, Wu. When we add this information to our program, it generates only the phylogenies that support these dialect groupings. Note that no existing phylogenetic system allows one to add such domain-specific information; in that sense, our method is novel.

With this preprocessed dataset and such domain-specific information, first we have computed the 3 phylogenies with the minimum number of incompatible characters. In each phylogeny, Yue is (Guangzhou), Hakka is (Liancheng, Meixian), Min is ((Taiwan, (Xiamen, Zhangping)), Fuzhou), Gan is (Anyi, Nanchang), Xiang is (Shuangfeng, Changsha), Mandarin is (((Beijing, (Yuci, Taiyuan)), Ningxia), (Chengdu, (Yingshan, Wuhan))), and Wu is ((Ningbo, (Suzhou, (Shangai 1, Shangai 2))), Wenzhou).

After that, for each phylogeny, we have computed the 294 perfect phylogenetic networks, with the minimum number of lateral edges. Only in 12 of these networks, the number of characters that require more than one lateral edge to become compatible is minimum; all of them are built on the phylogeny that groups the dialectal groups as follows: (Old Chinese, ((Yue, (Hakka, Min)), (Wu, (Gan, (Xiang, Mandarin))))). Two of these 12 networks are more plausible than the others, due to geographical distance, demography, etc. Both networks contain 3 contacts in Mandarin (pre-Beijing ↔ pre-Yingshang, pre-Beijing ↔ pre-Chengdu, pre-Beijing ↔ pre-proto-(Yingshan,Wuhan)), 1 contact in Min (pre-Zhangping ↔ pre-Fuzhou), and 4 contacts between dialectal groups.

The phylogeny above groups Chinese dialects into two: Southern dialects (Yue, Hakka, Min) and Northern dialects (Wu, Gan, Xiang, Mandarin). This sub-grouping is supported by linguistic evidence, e.g., isogloss, and other evidence, e.g., population movements. The grouping of Beijing as a sister to (Yuci, Taiyuan) coincides with results based on semantic innovations. However, they may tell an earlier story than the phonological studies do. In the two plausible phylogenetic networks described above, the contacts in Mandarin capture the influence of recent waves from Beijing Mandarin to Chengdu, Yingshan, and Wuhan. The contact in Min can be explained in terms of geographical distance, demography, self-identity, common religion, cultural traditions.