Assisted Reconstruction: the cases of Panoan and Mataco-Guiacuruan

There are many parallels between historical linguistics and molecular phylogenetics. We implemented an algorithmic pipeline that mimics, as closely as possible, the traditional workflow of language reconstruction known as the comparative method. The pipeline consists of suitably modified algorithms based on recent research in bioinformatics, that are adapted to the specifics of linguistic data. This approach can alleviate much of the laborious research needed to establish proof of historical relationships between languages. Equally important to our proposal is that each step in the workflow of the comparative method is implemented independently, so language specialists have the possibility to scrutinize intermediate results.

We have used our pipeline to investigate two language groups from South America, based on the lexical data from the Intercontinental Dictionary Series (IDS). The results of these tests show that the current approach is a viable and useful extension to historical linguistic research. Our first test set consists of four Panoan languages. For this data set, we investigated whether we could replicate accepted views about historical processes. The second test set was a set of seven Mataco-Guaicuruan languages. Here, we ask the question if we could find any evidence in the data supporting the hypothesis that Mataco and Guaicuruan form one language family.

As might have been expected, not much evidence is found linking the Mataco and Guiacuruan groups, although a few interesting lookalikes seem interesting enough to form a basis for further research. Most importantly, though, our test cases show that automatic methods can be used in tandem with manual interpretation to speed up historical comparison. We think such developments of assisted reconstruction are needed to improve our knowledge about the history of South American languages, which otherwise might take many decades to take place. With the growing body of available data, it also quickly is becoming too laborious to manually process all data for historical comparison. This kind of work seems to be an ideal topic of more extensive collaboration between linguists and computer scientists.