Internship proposal in Historical Computational Linguistic

The comparative method is a staple tool of historical linguistics. However, one of its main limitations, especially so when it comes to framing it as a natural language processing task is the lack of long time-scale gold data. While some scholars have estimated a capacity to backtrack 6000 to 8000 years worth of phonological changes for the comparative method, the earliest texts that can be read with any confidence are no more than 4000 years old and their partly logographic nature obfuscate a part of the actual phonological content. So that in practice it is impossible to test the earliest reconstructions against a ground truth.

The idea behind this project is to propose a solution to this limitation under the shape of artificially generated very long time-scale phonological histories. Artificially generated data have a number of strengths for testing computational approaches to the comparative method:

1. they can span much more than 4000 years, since we can generate arbitrarily long histories,
2. we can keep track of every intermediate state of each languages, thus we have access to much more information than is available for actual recorded languages even on the same time scale,
3. we can parameterize the evolution process, thus allowing us to test various hypothesis about both language evolution and the strength of different algorithmic methods.

The goal of the internship is to improve on an already existing system created during a previous internship. While the basics of phonological drifts have been set, there remain a number of processes to be fully implemented such as chained changes (p > b > v) or the apparition of tones, amongst other. The parametrization of changes seen as a stochastic process also needs further work. Then, depending on the length of the internship and the state of the generation system, we could imagine building on it to propose an actual reconstruction algorithm based on a reversed evolutionary process using Monte Carlo methods to search for possible past histories.

Candidates should be familiar with Python, and not scared of starting from already existing code. Basic git (pull, push, branches) is also recommended though not strictly necessary. They should also be interested in languages and their evolution.

If you are interested, please contact Mathieu Dehouck at : mathieu.dehouck [at] ens.psl.eu .