Quantitative and Formal Methods in Historical Language Comparison

Johann-Mattis List

Department of Linguistic and Cultural Evolution

Max Planck Institute for Evolutionary Anthropology

Table of Contents

Day 1: From Cognates to Correspondences

1.1 Introduction

Basic concepts of the course will be introduced in this session, and we will very briefly discuss questions of data representation, data modeling, and data analysis, and discuss the major objectives of computer-assisted as opposed to computer-based approaches to historical language comparison.

1.2 Cognates

We discuss the construct of cognacy by pointing to issues in the usage of the term and the application of the concept as reflected in the annotation of cognates in datasets or in the literature. As a central goal of the discussion, we want to raise awareness regarding the complexity of cognate relations, which have a direct impact on the praxis of historical language comparison.

1.3 Alignments

Alignments are introduced as the central concept that we need to use in order to identify regular sound correspondences and to prove that words are truly cognate. While so far only occasionally being used by practictioners of classical approaches to historical language comparison, we will try to advertise the usage of alignment analyses both in prose texts as well as in the construction of large comparative datasets.

1.4 Correspondences

Based on the notion of phonetic alignments, we will take a closer look at regular sound correspondences, and how corresponcences can be assembled to form patterns from aligned cognate sets. The notion of sound correspondence patterns will be crucial later when discussing formal approaches to phonological reconstruction.

Day 2: From Words to Trees

2.1 Lexical Motivation

Complex words are frequent in the languages of the world and often also occur in the basic vocabulary of human languages. At times they may even provide important evidence for linguistic reconstruction. In order to handle complex words properly, we need find ways to handle them formally. Initial ideas regarding the handling of complex words with the help of morpheme glosses are introduced and their implications are discussed.

2.2 Phylogenies

We use the term *phylogeny* to refer to evolutionary scenarios that describe how languages evolved into their attested shape. Phylogenies are not restricted to phylogenetic trees, but can often be represented by one or more trees at the same time. We discuss the concept of phylogenetic trees, of phylogenetic networks, and the difficulties of reconciling evolutionary scenarios for individual words with overarching language phylogenies.

2.3 Distance-Based Approaches to Phylogenetic Reconstruction

As a starting point for the exploration of a phylogenetic datasets, distance-based approaches provide a quick summary on the data we can observe. We introduce basic concepts and discuss two major algorithms for the reconstruction of phylogenetic trees with the help of distance-based approaches.

2.4 Character-Based Approaches to Phylogenetic Reconstruction

Character-based approaches to phylogenetic reconstruction model the evolution of each individual character (usually a cognate set in the data) and try to find an overarching phylogeny that explains individual evolutionary patterns best. We discuss the advantage of character-based over distance-based approaches and give a brief introduction to parsimony-based phylogenetic reconstruction as well as a very brief overview about Maximum Likelihood, the core concept underlying the nowadays very common Bayesian approaches to phylogenetic reconstruction.

Day 3: From Words to Stars

3.1 Introduction

Discuss major aspects of phonological reconstruction as the core principle of linguistic reconstruction.

3.2 Reconstruction without Trees

Discuss the — in our opinion — false claim that scholars reconstruct with a tree in mind. Show, instead, that concrete ideas about common sound change processes are more likely to drive reconstruction choices, driven by formulaic principles by which proto-forms for unique sound correspondence patterns are identified.

3.3 Supervised Phonological Reconstruction

Preset a recent framework that learns how to reconstruct based on phonetic alignments and sound correspondence patterns and show how this could be useful for ongoing work in the field of comparative linguistics.

3.4 Uncertainty in Phonological Reconstruction

Present recent ideas on handling uncertainty in phonological reconstruction, inspired by automated methods for phonological reconstruction.

Day 4: From Words to Deeds

4.1 Overview

Provide a rough overview about available tools that can be used in quantitative approaches to historical language comparison. Discuss major data formats that are typically needed, when carrying out an analysis.

4.2 LingPy

Give a short introduction into the major features of the LingPy Python library and how it can be used to analyze one's data.

4.3 EDICTOR

Give a short demo of the EDICTOR tool for the creation and curation of etymological dictionaries in a consistent, and formal framework that makes use of alignments, correspondence patterns, and allows for a rigorous checking of one's data.