

*Polymorphic characters in Indo-European languages*

Donald Ringe, Tandy Warnow, and Steve Evans

The estimation of evolutionary history in biology is a fundamental part of much biological research, with long-established techniques going back for more than half a century. Many of the early evolutionary trees established in biology were based upon “morphology”. However, the increasing availability of DNA sequence data, and the introduction of stochastic models of DNA sequence evolution, dramatically changed molecular systematics. Not only did the methods for inferring evolution change as a result of the new data and models, but the estimations of history also were modified, sometimes in significant ways. While much still needs to be understood, without doubt the viewpoint of evolution as a stochastic process has itself brought about a fundamental change in the way molecular phylogenetics is done.

Our recent research has been focusing on developing appropriate stochastic models of character evolution in languages, in the hope that statistical methods for estimating language evolution based upon these models will be able to address some of the major open problems in the history of the Indo-European and other language families. In our earlier work, we postulated a simple model of language evolution. This model described how characters (lexical, morphological, and phonological) evolved, both through “genetic transmission” and through borrowing between language groups; our model also allowed for limited (and identifiable) homoplasy, where “homoplasy” refers to the repeated appearance of a character state in the evolutionary history, either through back-mutation or parallel evolution. This model seems reasonable for most characters, but makes the critical assumption that characters are always “monomorphic”, which means that each character has a single state on any given language. By contrast, characters with two or more states on a given language are called “polymorphic”. A simple example of a polymorphic character in linguistics is the semantic slot ‘small’ in English, as both *little* and *small* have the same basic meaning. While the frequency with which linguistic characters violate the assumption of monomorphism is not yet known, our analysis of the Indo-European family turned up many polymorphic characters. A consideration of the various methods currently used to reconstruct evolution in languages makes it immediately clear that none of these can reasonably utilize polymorphic characters.

This paper will focus on the problem of polymorphic characters using examples from Indo-European, since thorough traditional work on that family has yielded a large body of useful facts. As we have established in previous work, polymorphism is chiefly a property of lexical characters, and such characters must be defined semantically in order to render them informative. Polymorphism typically arises by the process of semantic shift, in which a word of meaning ‘x’ begins to acquire a new meaning ‘y’, competing with an already existing word that has that meaning. Two properties of this process render polymorphic characters difficult to use for cladistic purposes in the absence of a good stochastic model of language evolution. On the one hand, some semantic shifts are very common, recurring frequently in historically unconnected lineages, while others are much rarer. On the other hand, the polymorphism created by semantic shift is sometimes transient (in which case it may escape the historical record) and sometimes persistent; an instance of the latter is the English case adduced above, which has persisted for at least a millennium (see the entries for *little* and *small* in the Oxford English Dictionary). Since the borrowing of vocabulary from language to language also involves a lateral transfer of linguistic material into a semantic slot which defines a lexical character, it can also give rise to polymorphism. The factors governing the frequency of borrowing are quite different from those that govern semantic shift, but once a polymorphism has arisen through borrowing it can behave like any other.

We will propose some stochastic models of character evolution which produce different patterns of polymorphism and consider the problem of estimating evolutionary histories under these models.