

Refining variant statements for phylogenetic analysis and building a community of interest.

This paper will address two issues. The first is a matter of long-standing difficulty relating to the expression of variation among texts, in order to achieve both optimal presentation and analysis of the differences among them. So far, most variation among witnesses submitted to phylogenetic analysis has been expressed in what is called the 'parallel segmentation' method. This aligns all variants across all witnesses at a place of variation, so that any one witness can be compared directly with any other. This has the virtue of simplicity, and provides readers with a transparent way of seeing how any one witness relates to any other. Records of variation prepared in this way have also given good results when submitted to phylogenetic analysis.

However, it has been long recognized that this method of expression of variation is less than optimal, both in terms of seeing exactly how witnesses relate to each other and for purposes of phylogenetic analysis. This is particularly the case with transposition: in (say) a six-word phrase, if one manuscript among (say) fifty moves the first word to the sixth word, then the parallel segmentation method means that all fifty manuscripts are measured against the same six word variant segment. This means that finer-grained variation and relationship among the fifty manuscripts, which might differ only by one word or two among the six words, is quite lost. In consequence, too, all variants on this six-word segment seem equally distant from each other. This leads to a considerable loss of information, as some manuscripts among the fifty will be much closer to each other than are others, and it would be useful both for the human reader to see this clearly, and for the phylogenetic processing to have this input. In an important paper, Matt Spencer and Chris Howe outlined a solution to the problem, based on 'multiple progressive alignment'. However, they did not propose a formal method of expressing the results of this alignment in a manner efficient both for scholars and for phylogeneticists. A formal expression would be useful as it could then become a target for different methods of achieving multiple progressive alignment. This is particularly timely as there are currently several initiatives seeking to create new (and better) collation systems.

The second issue this paper will address is the need to build a community of interest based around the application of quantitative methods of analysis (principally, but not exclusively, phylogenetics) of relations in textual traditions. The current set of workshops is a crucial step towards this, and those attending the Helsinki meeting (and, we hope, its successors) may form the nucleus of this community. I will discuss some ways we can foster the development of this community, and I will announce the launch of a website dedicated to phylogenetic data.