Erich Round (University of Queensland)

Dataset design processes need to be scientifically reported: The sensitivity of Bayesian clustering and 'researcher degrees of freedom'

Modern computational methods open up new directions of enquiry utilizing cross-linguistic datasets whose size would overwhelm traditional pencil-and-paper analyses. Naturally however, questions arise when using techniques not originally designed for linguistic research, as to how performance might be affected by properties of linguistic datasets, whose principles of design have yet to alter much from the paper-and-pencil era. Issues such as non-independence of variables may be innocuous in some cases (Pagel & Meade 2006), but precisely when and whether this generalizes requires ongoing investigation.

In a series of experiments we examine the performance of the Bayesian clustering algorithm STRUCTURE (Pritchard et al. 2000) by making controlled variations to <10% of a large cross- linguistic dataset of typological variables (Reesink et al. 2009—121 languages; 155 binary variables; 88% of cells complete). We find that in terms of both the number of clusters inferred, and the posterior probability of those inferences, STRUCTURE is remarkably sensitive to the inclusion/exclusion of variables which have:

i. A high proportion of the same value across languages;

ii. A high number of missing values;

iii. A tendency to pair-wise implication with another variable (e.g. A=1 very often entails B=0);

iv. A high pair-wise correlation with another variable

This may carry consequences for the methodology of modern linguistic dataset construction, since a typologist's decision over whether or not to build variables A, B or C into a dataset may hinge precisely on issues such as (i–iv), which can then influence the final statistical result — a manifestation of the "researcher degrees of freedom" problem (Simmons et al. 2011). Thus, to promote transparency and aid evaluation of research, we advocate that when datasets are designed, researchers explicitly document and publish the accompanying decision-making process, since as our results indicate, such decisions could bear significantly upon eventual results, before language coding even begins.

Pagel M., Meade A. (2006) Estimating rates of lexical replacement on phylogenetic trees of languages. In: Forster P, Renfrew C, editors, *Phylogenetic methods and the prehistory of languages*, McDonald Institute for Archaeological Research: UK, McDonald Institute Monographs. 173–182.

Pritchard J., Stephens M., Donnelly P. (2000) Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.

Reesink G., Singer R., Dunn M. (2009) Explaining the linguistic diversity of Sahul using population models. *PLoS Biology* 7, e1000241.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366.