Towards a singular measure of varietal similarity:  
The Structural Concordance Criterion

The overarching pursuit of much work in comparative sociolinguistics is the question of whether—or the degree to which—varieties are related. The present work argues for two methodological advances in this pursuit, where such additional analytic and statistical methods enable both a more formal and potentially more illustrative view of variation.

Much work in comparative sociolinguistics is informed by the hypothesis that “underlying grammatical structure can be discerned from examination of the distribution and conditioning of synchronic variability” (Poplack & Tagliamonte, 1991:331). This often involves fitting regression models separately to data from two or more varieties, then making side-by-side comparisons of factor weights (Varbrul/Rbrul) or coefficients (R) from these separate regressions. Excellent comparative work has been built upon this approach, yet it has substantial limitations. Differing sample sizes make comparative discussion of what comes up as significant suspect; varying N affects what is retained in stepwise selection procedures. Next, collinearity among factors can result in different factors being selected for different varieties. Most importantly, judgment of what constitutes “similar” across separate regressions is necessarily informal. While statistical significance within each regression is tested, comparisons across regressions are typically not.

The appropriate formal method we argue for involves interaction analysis. Where possible—and in truth this is not always practical—the data from multiple varieties can be gathered into a single large regression, with source variety itself encoded as a factor. We can now formally examine the statistical interaction between source variety and any given factor. This offers a testable statistical measure of weights across varieties. In the present work we illustrate with data from studies of American and Caribbean creole English varieties (Rickford 2010), including cases where factor differences that might seem substantial on visual inspection turn out less than statistically significant on formal analysis, as well as cases where a rather uncertain judgment is in fact confirmed as significant variation.

Finally, however, even formal interaction analysis proceeds by individual constraint. We lack any single statistical measure of overall similarity/dissimilarity among varieties. Here we propose a new metric, the Structural Concordance Criterion (SCC), as a supplement to the narrower individual constraint view. Formally, SCC is calculated as

\[
(\hat{\beta}_1 - \beta_r)(X'X)(\hat{\beta}_1 - \beta_r) + (\hat{\beta}_2 - \beta_r)(X'X)(\hat{\beta}_2 - \beta_r)
\]
where $\hat{\beta}_i$ is an estimated Beta coefficient (akin to factor weight) from a given separate regression. Structurally, this is not dissimilar to correlation calculation, though here we allow both for varying sample sizes $n_i$, with “pooled” Beta coefficient $\beta_P$ corrected as

$$\beta_P = \frac{n_1\hat{\beta}_1 + n_2\hat{\beta}_2}{n_1 + n_2}$$

and account for potential collinearity among factors via matrix transformation. Once again, the present work illustrates SCC in comparing American and Caribbean creole English varieties.

Certainly, no statistic can or should remove the scientist from the comparison process, and as with any measure, thoughtful interpretation in appropriate context remains important, but use of SCC offers a potentially useful additional tool in the comparative sociolinguist’s tool chest.

**References**
