The Delta Spreadsheet

David Hoover *(david.hoover@nyu.edu)*
New York University

John F. Burrows introduced Delta, a simple measure of authorial difference in his Busa Award lecture (2001), and further elaborated upon it in three articles (2002a, 2002b, 2003). In all of these discussions Burrows relies on an Excel spreadsheet that helps to simplify and partially automate the calculation of Delta. At the *ALLC/ACH* conference in Gothenburg, David L. Hoover presented the results of further tests of Delta on prose and discussed a more complex version of Burrows's spreadsheet that takes the automation of the calculation and the analysis of results much further (2004a), and he has just published two articles that rely on such spreadsheets (2004b, 2004c).

Given the burst of activity in authorship attribution circles following the introduction of Delta, many researchers are interested in using it on various projects. Unfortunately, even Hoover's 2004 versions of the spreadsheet are rather daunting in their complexity, and their macros are difficult to understand because they do not include comments. Further, the researcher must do substantial analytical work on raw word frequency lists before they can be inserted in the spreadsheet for Delta testing. Once the lists are produced, the frequencies must be transformed into text percentages and a zero frequency record must be inserted in the list for each text if any of the most frequent words does not occur in that text. This is not a significant problem for analyses using only a small number of the most frequent words because nearly all of them will occur in each text, but, as Hoover has shown (2004b, 2004c), increasing the word list to the 700-800 most frequent often improves the accuracy of the analysis, and many of the 800 most frequent words will normally fail to appear in one or more of the texts. Manually adding zero records may be an acceptable method in small analyses, but it would be an extremely time-consuming and error-prone process if the 800 most frequent words in a set of fifty or more texts were to be analyzed.

Hoover's analyses also show that removing personal pronouns and words for which a single text provides nearly all the occurrences significantly improves Delta (and other kinds of statistical analyses of authorship), and these are non-trivial processes that are difficult enough to prevent some researchers from trying out these techniques. The addition of the various possibilities for *Delta Prime* introduced in Hoover's second article (2004c) makes for still greater complication, and seems likely to prevent the interested humanist who is not an Excel maven from further testing these innovative measures on new corpora and from using them in real authorship attribution problems.

My current project involves further elaboration of Hoover's spreadsheets to automate more of the necessary processes. Beginning with a version provided by Hoover that includes explanatory comments on the macros by Marc LeBlanc of Wheaton College (MA), I hope to produce a spreadsheet that can accept as input a list of the authors and texts, the raw word frequencies from the corpus as a whole and from the individual primary and test texts. The complete analysis will be performed within the spreadsheet itself. This will allow anyone who has access to any of the myriad of software tools that can produce ranked frequency lists to try out Delta and the various Delta Primes without needing to have expertise in text analysis, Excel, or Visual Basic. The project is currently under way, with the various formulas for calculating Delta and the various Delta Primes already added and the analytic work planned out and in progress. Initial testing has begun to determine whether or not the macros will operate with acceptable speed, and whether the limitations of Excel will impact the number of frequent words that can be analyzed. If performance proves too poor, I intend to use other methods than Visual Basic and link them as seamlessly as possible with the spreadsheet. By the time of the conference, I expect to have a fully operational version to demonstrate and distribute to anyone who is interested.

A secondary benefit of the current project is more wide ranging, having to do with the question of how to balance using the good tools for performing the analysis and manipulation of the word frequency lists (certainly Visual Basic is not one of them!), and providing a tool that is usable by the largest possible number of users, even if those users are not particularly computer literate. This has long been a question of serious interest to software developers, and the relatively small scale of this project may allow it to come to the fore in interesting ways. I hope to benefit from the expertise of conference attendees in continuing to develop and improve The Delta Spreadsheet.

Bibliography

Burrows, J.F. "'Delta': a measure of stylistic difference and a guide to likely authorship." *Literary and Linguistic Computing* 17 (2002a): 267-287.

Burrows, J.F. "The Englishing of Juvenal: computational stylistics and translated texts." *Style* 36 (2002b): 677-99.

Burrows, J.F. "Questions of Authorship: Attribution and Beyond." *Computers and the Humanities* 37.1 (2003): 5-32.

Burrows, J.F. "Questions of Authorship: Attribution and Beyond." Paper delivered at the Association for Computers and the Humanities and Association for Literary and Linguistic Computing, Joint International Conference. June 14, 2001.

Hoover, D.L. "Testing Burrows's Delta." Paper delivered at the Association for Literary and Linguistic Computing and Association for Computers and the Humanities, Joint International Conference, Göteborg, Sweden. 2004a.

Hoover, D.L. "Testing Burrows's Delta." *Literary and Linguistic Computing* 19.4 (2004b): 453-475.

Hoover, D.L. "Delta Prime?" *Literary and Linguistic Computing* 19.4 (2004c): 477-495.