
Updating Delta and Delta Prime

David L. Hoover (david.hoover@nyu.edu)

New York University

John F. Burrows's Delta, a unitary measure of textual difference, has created a flurry of interest in authorship attribution and statistical stylistics since its introduction in 2001 (Burrows 2001, 2002a, 2002b, 2003, 2005; Hoover 2004b; García and Martín 2006; other studies are in progress). Burrows uses a Microsoft Excel spreadsheet to simplify and partially automate his calculation of Delta, and I have presented increasingly complex versions of The Delta Spreadsheet that more fully automate the calculation and the analysis of results (2004a, 2005a). I have also suggested possible improvements in how Delta is calculated or defined, which I have tentatively called "Delta Primes" (Hoover 2004c, 2005a, 2006).

The growing popularity of Delta analysis will surely lead to its further increased use. In addition to a number of studies in progress in the broad area of humanities computing, studies by other researchers are either already ongoing or being planned using Delta in the evaluation of high-stakes educational testing of writing in high schools, in the "authorship" of film direction, in evolutionary biology, and in studies of adverse drug reactions. The simplicity of automated versions of Delta analysis is critical to its increased use, especially for those who are not specialists in authorship attribution or humanities computing.

In its current state, The Delta Spreadsheet allows the user to paste in raw word frequency lists and then to perform a complete Delta analysis automatically by running a Visual Basic macro that calls other macros. One of the macros removes any words from the word list that are not found in the primary set of texts (the presence of such words would prevent the calculation of Delta by causing division by zero). Another macro changes the raw word frequencies into text percentages and inserts a zero frequency record in the word list for each text whenever any of the most frequent words does not occur in that text. This process is extremely tedious, time-consuming, and liable to error if done manually, especially on the large word lists that are now typically used in Delta analysis (800-4,000 words). Another optional macro removes personal pronouns. These pronouns are entered into a column in the spreadsheet, along with the master word list to be analyzed, and the user also can enter there any other words that should be eliminated (for example, noise words or proper names). The spreadsheet also allows the user to specify whether the word list should be culled to remove words for which a single text provides most of the occurrences,

and, if so, to specify what percentage of occurrences should be used as the cut-off. The user can also specify the size of the word list to create for the analysis and the total number of words to analyze; the analysis macro can also be set to run several times based on increasing or decreasing numbers of words. The result of all this automation is that an entire Delta analysis can be performed as a background task. And, on modest numbers of texts that are not very large, an entire analysis can be performed easily in an hour. This allows a researcher to try many different combinations of options in the search for the most accurate and reliable results.

My current project involves further elaboration of The Delta Spreadsheet to automate more of the necessary processes. I have created an additional spreadsheet into which the user can enter a list of texts to be processed. Once the texts and their authors have been entered and the primary and secondary text sets specified, the user runs a macro that collects the word frequency lists of these files from the current directory and adds the appropriate text and author labels. This allows the user then to cut and paste the word lists into The Delta Spreadsheet for processing, which simplifies the process, saves time, and reduces error. A third spreadsheet designed for quicker analysis of results allows the user to paste in the results of a series of Delta analyses and run a macro that reformats, sorts, and prepares the data for graphing, with similar benefits.

A final significant upgrade of The Delta Spreadsheet takes advantage of an analysis by Shlomo Argamon (forthcoming) of the statistical bases of Delta. Argamon shows that Delta (as well as my proposed Delta Primes) can be calculated without relying on the mean frequencies of words in the primary set of texts. The revised method of calculation makes possible a more streamlined version of The Delta Spreadsheet that allows me to increase the number of texts that the spreadsheet can process. It should also improve the performance of the macros, which can take several hours to process a large number of long novels. My poster will show how the modifications improve the performance of the spreadsheets and will present an example with a very large number of authors. I will also be prepared to do a live software demonstration of the operation of the spreadsheets at the conference, both on already existing word frequency lists and on texts which conference attendees supply.

Bibliography

Argamon, Shlomo. "Interpreting Burrows's Delta: Geometric and Probabilistic Considerations." *Literary & Linguistic Computing* (Forthcoming).

Burrows, John F. "Questions of Authorship: Attribution and Beyond." Paper presented at ALLC/ACH Joint International Conference, New York, June 14, 2001. 2001.

Burrows, John F. "Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary & Linguistic Computing* 17.3 (2002a): 267-287.

Burrows, John F. "The Englishing of Juvenal: Computational Stylistics and Translated Texts." *Style* 36 (2002b): 677-99.

Burrows, John F. "Questions of Authorship: Attribution and Beyond." *Computers and the Humanities* 37 (2003): 5-32.

Burrows, John F. "Who Wrote Shamela? Verifying the Authorship of a Parodic Text." *Literary & Linguistic Computing* 20.4 (2005): 437-450.

Burrows, John F. "All the Way Through: Testing for Authorship in Different Frequency Strata." *Literary & Linguistic Computing* (2006). Advanced Access published January 6, 2006

Garcia, A. M., and J.C. Martin. "Function Words in Authorship Attribution Studies." *Literary & Linguistic Computing* (2006). Advance Access published November 14, 2006

Hoover, David L. "Testing Burrows's Delta." Paper presented at ALLC/ACH Joint International Conference, Göteborg, Sweden, June 11-16, 2004.

Hoover, David L. "Testing Burrow's Delta." *Literary & Linguistic Computing* 19.4 (2004a): 453-475.

Hoover, David L. "Delta Prime?" *Literary & Linguistic Computing* 19.4 (2004b): 477-495.

Hoover, David L. "Delta, Delta Prime, and Modern American Poetry: Authorship Attribution Theory and Method." *ACH/ALL 2005 Conference Abstracts*. Victoria: University of Victoria Humanities Computing and Media Centre, 2005a. 83-84.

Hoover, David L. "The Delta Spreadsheet." *ACH/ALLC 2005 Conference Abstracts*. Victoria: University of Victoria Humanities Computing and Media Centre, 2005b. 85-86.

Hoover, David L. *The Delta Calculation Spreadsheet Online*. 2005c. <<http://www.nyu.edu/gsas/dept/englis h/dlh/TheDeltaSpreadsheets.html>>

Hoover, David L. "Word Frequency, Statistical Stylistics, and Authorship Attribution." *Advanced ICT Methods Guide to Linguistics*. Ed. T. McEnery. Forthcoming.