The Craig Zeta Spreadsheet

Hoover, David L.
david.hoover@nyu.edu
New York University, USA

Zeta, a new measure of textual difference introduced by John F. Burrows, can be used effectively in authorship attribution and stylistic studies to locate an author's characteristic vocabulary–"marker" words that one author uses consistently, but another author or authors use much less frequently or not at all (Burrows 2005, 2006; Hoover 2007a, 2007b, 2008). Zeta analysis excludes the extremely common words that have traditionally been the focus and concentrates on the middle of the word frequency spectrum. Beginning in 2008, I developed an Excel spreadsheet implementation of Zeta and its related measure Iota (which focuses on words that are relatively rare in one author, but extremely rare or absent from others), available on my Excel Text-Analysis Pages. Recently, however, Craig has developed an alternative version of Zeta that simultaneously creates sets of marker words and anti-marker words and has applied it impressively to Shakespeare authorship problems (Craig and Kinney, 2009; Hoover, forthcoming). Although Craig's Zeta focuses on the same part of the word frequency spectrum as Zeta, its calculation and results are quite different. Because it seems poised to become an important tool for computational stylistics, I have created and will demonstrate the Craig Zeta Excel spreadsheet that automates its calculation.

Craig's Zeta is a powerful but simple method of measuring differences among authors. It begins with two sets of texts divided into about equal-sized sections, then compares how many sections for each author contain each word, ignoring the frequencies of the words and concentrating on their consistency of appearance. The most natural comparison is between two authors, but it can be used to study any contrast. My demo version contrasts thirteen female and thirteen male American poets born between 1911 and 1943 (about 8,000 words of poetry by each, divided into two sections).

The heart of the method is that it combines the ratio of the sections by one author in which each word occurs with the ratio of the sections by the other author from which it absent into a single measure of distinctiveness for each word. Zeta scores theoretically range from two (for a word found in every section by one author and absent from every section by the other), to zero (for a word found in no sections by one author and in all sections by the other). Sorting the words on this composite score produces two lists of words, one favored by the first author and avoided by the second, the other favored by the second author and avoided by the first.

The snippet from the spreadsheet in Fig. 1 (shown before the macro operates) will clarify the calculation. In E7 and E8, the user enters labels (automatically copied into columns A and G and Row 9) for the two groups to be compared. The data to be analyzed is in columns H through CA, rows 11ff, with most columns minimized so the various categories are visible. The combined word frequency list for the two groups is in column G, with the raw frequencies for each word in each section in columns H-CA. The calculation is performed in columns A-E. Column D sums the sections of poetry by women that contain the word, and column E sums the sections of poetry by men that do not contain the word. The most frequent words typically occur in all sections, but note that *me*, the 30th most frequent word, is absent from one of the men's sections. Column B calculates the ratio of women's sections containing the word to the total number of women's sections; column C calculates the ratio of men's sections not containing the word to the total number men's sections. Column A sums columns B and C to produce the Zeta scores. Columns H-CA of row 1 show the number of different words (word types) in each section, and below them, the percentage of types that are marker words for women or men (these are not meaningful until the macro has operated). In cells F2-F3 the user can set the number of marker words for each group at different levels for sections of different sizes and can see three sets of results at once in H-CA.

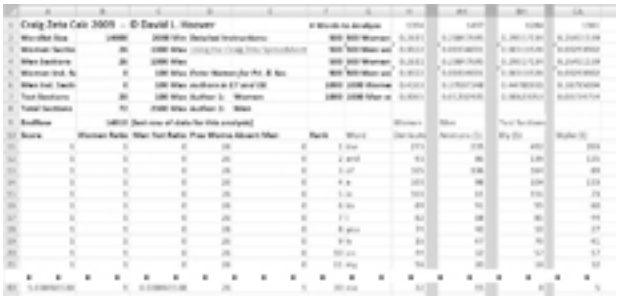

Figure 1

An Excel macro automates the calculation of Zeta. Word frequency lists for the texts to be analyzed are entered into five sub-sheets (not shown). One contains the sections of the primary group and one contains the sections of the secondary group. Two more sub-sheets contain any independent sections by the primary and secondary groups (these can be used to test the method's success on known texts). Finally, there is a sub-sheet for the texts to be tested. The macro clears out old data, enters formulas into columns H-CA, rows 2-7, copies the texts out of the sub-sheets into the main sheet, shrinks the columns, and enters ranks for the words in column F. It then sorts the words on their Zeta scores in column A, descending, so that the words most distinctively used by women appear at the top and those most distinctively used by men are at the bottom. It selects the 1000 most distinctive men's words and resorts them in reverse order, with the most distinctive at the top of their section. The sheet can handle 15,000 words, but the sample above analyzes 14,000 words (calculated in cell B2), so rows 11-1010 will contain the 1000 most distinctive women's words and rows 13,011-14,010 the 1000 most distinctive men's words.

Figure 2 shows the data after the macro runs. Here *mother's*, found in 13 of 26 sections by women and absent from 24 of 26 sections by men, is the most distinctive women's word in this comparison. The most distinctive men's word, *cross*, is found in just 3 of 26 sections by women, but is absent from only 10 of 26 sections by men. The most distinctive words for each group are found in columns CB and CC in descending order of distinctiveness. The figures in rows 2-7 of columns H-CA show how these distinctive words are distributed in each section. For example, H2-H3 shows that almost 15% of the words Derricote uses in this section are among the 500 most distinctive women's words, but only about 5% are among the 500 most distinctive men'swords. For Ammons (1), in AH2-AH3, the first section by a man, the proportions are roughly reversed.

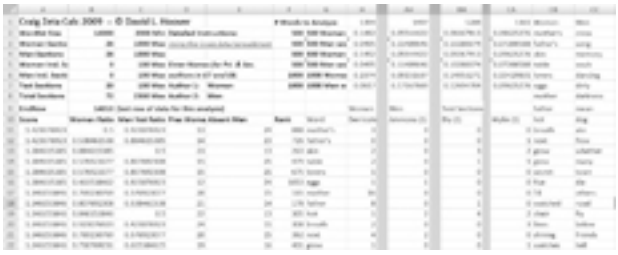

Figure 2

Columns BH-CA contain the same information for the test sections, 25 sections of poetry by seven female and seven male poets whose work had no part in the creation of the word list on which the analysis is based. Finally, the macro creates a scatter graph of the analysis with section names as labels (Fig. 3). The vertical axis records the proportion of types in each text that are marker words for men and the horizontal axis does the same for marker words for women. As Fig. 3 shows, twenty of the twenty-five new sections of poetry (80%) trend toward the group that matches their gender. The fact that these same words produce a similar, though slightly less accurate, result for seven male and seven female contemporary novelists is further evidence that the method is capturing some kind of genuine difference. This is a strong result, especially given the relatively small samples and the wide diversity that characterizes 20th century American poetry. The little chart that follows Fig. 3 shows some of the clusters of related words that occur in the women's and men's marker words, and hints at the kinds of analysis that Craig's Zeta makes possible.

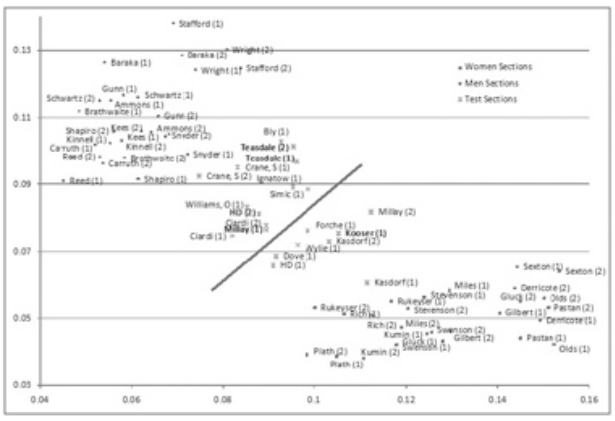

Figure 3

Cluster	*500 Women's Markers*	*500 Men's Markers*
Family	*mother's, father's, mother, father, children, ancestral, aunt, baby, birth, child, child's, cousins, daughters, family, generations, uncles*	
Religion	*altar*, *nuns*, and *praying*	*faith, heaven, hell, prayers, souls, spirit, Christ, gods, myth, paradise, religion, spirits, temple*
Houses/ Furniture	*danced*	*song, dancing, sing, dance, sang, dancer, music, singer, singing, sings*
Personal Pronouns	*he'll, I'd, mine, ourselves, she'd, she's, they'd, you'd, you're, yourself*	

References

Burrows, J. F. (2006). 'All the Way Through: Testing for Authorship in Different Frequency Strata'. *LLC.* **22**: 27-47.

Burrows, J. F. (2005). 'Who wrote Shamela? Verifying the Authorship of a Parodic Text'. *LLC.* **20**: 437-450.

Craig, H., and Kinney, A., eds. (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

Hoover, D. (2007a). 'Corpus Stylistics, Stylometry, and the Styles of Henry James'. *Style*. **41**: 174-203.

Hoover, D. (2007b). 'Quantitative Analysis and Literary Studies'. *A Companion to Digital Literary Studies*. Susan Schreibman, Ray Siemens (eds.). Oxford: Blackwell, pp. 517-33.

Hoover, D. (2008). 'Searching for Style in Modern American Poetry'. *Directions in Empirical Literary Studies: Essays in Honor of Willie van Peer*. Sonia Zyngier, et. al. (eds.). Amsterdam: John Benjamins, pp. 211-27.

Hoover, D. (2009). 'The Excel Text-Analysis Pages'. `https://files.nyu.edu/dh3/public/The%20Excel%20Text-Analysis%20Pages.html`.

Hoover, D. (2010: forthcoming). 'Authorial Style'.