WORD FREQUENCY AND KEYWORD EXTRACTION
AHRC ICT Methods Network Expert Seminar on Linguistics
8 September 2006, Lancaster University, UK

Word Frequency, Statistical Stylistics, and Authorship Attribution

David L. Hoover, New York University

Keywords
word frequency, authorship, statistical stylistics, style, corpora, Victorian novels, contemporary American poetry.

Abstract
The availability of large corpora and of electronic texts has renewed interest in the venerable topic of word frequency. Innovations in analytic techniques and in the ways word frequencies are selected for analysis have also been instrumental. Authorship attribution and statistical stylistics have, until recently, typically been based upon fewer than the 100 most frequent words of a corpus. These words – almost exclusively function words – are attractive because they are so frequent that they account for most of the running words of a text, and because such words have been assumed to be especially resistant to intentional manipulation by an author, so that their frequencies should reveal authorial habits that remain relatively constant across a variety of texts.

Recent work on style variation has suggested that selecting words because of their frequency in sections of texts rather than in the entire corpus is more effective in capturing stylistic shifts. Removing words that are frequent overall only because they are very frequent in a single text has also been shown to dramatically improve the accuracy of an analysis. Another trend has been to increase the number of words analyzed to as many as the 6000 most frequent words, a point at which almost all the words of the text are included and almost all are content words. Finally, following the innovative work of John Burrows, there is a great deal of current interest in Delta, a new measure of the differences between texts that is based upon comparing how different the texts are from the mean for the entire corpus. Further refinements in the selection of words for analysis and in alternative formulas for calculating Delta suggest that further improvements in the accuracy may be possible and that we may be nearing a theoretical explanation of how and why word frequency analysis is able to capture authorship and style.

I will be discussing these issues mainly with reference to a 2,000,000 word corpus of contemporary American poetry and a much larger corpus of 46 Victorian novels.

The increased availability of large corpora and electronic texts, and innovations in analytic techniques have spurred a great deal of recent interest in the venerable topic of word frequency, especially in authorship attribution and statistical stylistics. Until recently, research in these areas has typically been based upon the 30 to 100 MFW (most frequent words) of a corpus. These words–almost exclusively function words–were chosen because they are so frequent that they account for a large proportion of the running words (tokens) of a text, and because of the assumption that their frequencies should be especially resistant to intentional authorial manipulation, and so should reveal authorial habits that remain relatively constant across a variety of texts. Successful as studies based on these words have been, recent work has expanded the list, changed the way the words for analysis are chosen, and proposed new analytic tests and measures of authorial style.

A discussion of some of these recent developments can usefully begin with an examination of the sixty MFW of Edith Wharton's *The Age of Innocence* (1920; *Age*, below), shown in Table 1. This typical novel has a total of 101,840 tokens and 9731 types, and roughly half its types (4873) are

hapax legomena (words occurring once). The rapid decrease in word frequencies shown in Fig. 1 is typical of English texts, as are the words themselves, though the feminine personal pronouns are more frequent than usual. Anyone working with word frequency lists will find this one familiar and very similar to those from corpora such as Brown and BNC. Although Brown is ten times and BNC a thousand times as large as this novel (and the latter British rather than American), forty-seven of these sixty words are also among the sixty MFW of both corpora. The presence of two proper names, *Archer* and *Oleska*, is also typical. (Any number of proper names from zero to about five could be considered typical.) Aside from *Archer* and *Oleska*, only *new* and *said* might be considered content words (*like* would drop in frequency if verb examples were subtracted), and *new* appears here only because it is so frequent as part of *New York.* It seems surprising that comparing such lists of ubiquitous and very frequent words is so regularly effective in distinguishing one author from another.

However, even an ordinary list like this one raises significant questions. Deleting the proper nouns before comparing this text with other texts seems appropriate, and many researchers have done so, sometimes without comment, but common nouns that might also be proper nouns remain problematic, and some are far more difficult to identify and deal with than *new*. Consider *woman* in William Golding's *The Inheritors*, a novel told mainly from the point of view of a Neanderthal as his society is destroyed by a more advanced invading tribe. *Woman* is more than fifteen times as frequent in *The Inheritors* (rank: 43) as it is in the written portion of BNC (rank: 387), primarily because of the frequency of 'the old woman' for the Neanderthal matriarch, and 'the fat woman' and 'the crumpled woman' for women of the invading tribe. Even without these epithets, however, *woman* ranks 221st, and remains about 2.7 times as frequent in *Age* as in BNC.

Such words raise subtle and potentially important analytic and theoretical questions. Is it possible, practical, or even desirable to tease out the different functions and meanings of *woman* in *The Inheritors*? How does this question relate to the thematic importance of *woman* in the novel? Are epithets, or proper nouns themselves, stylistic or authorial markers? Imagine *The Wizard of Oz* with *Dorothy Gale* and *Toto* replaced by *Tiffany Spindrift* and *Fifi*. And is the relationship between *Gale* and *tornado* irrelevant? Although the names *Archer* and *Oleska* are unusual enough that they are unlikely to occur in any novels being compared with *Age*, they certainly could do so, and names like *John*, *London*, or *New York* could potentially skew results in novels that were otherwise similar. In an analysis of only the sixty MWF, the absence of *Archer* and *Oleska* from other novels by Wharton would also tend to separate the novels in spite of their common authorship. Yet assuming their absence is unwise; some of Faulkner's characters, for example, appear in more than one of his novels, and *Archer* is also frequent in James's *The Portrait of a Lady*, in which Isabel Archer is the main character.

The problem of *new* as a common noun and as part of *New York* also suggests that POS (part of speech) tagging might be desirable to prevent one text in which *new* is extremely frequent as a common noun from appearing similar to another that happens to be set in New York. Unfortunately, even the most accurate POS taggers introduce errors, and all the taggers I have tried are hopelessly inaccurate for poetry. When, as is often true of my own work, newly-constructed corpora of millions of words are being analysed, manual correction of tagging is impractical. Furthermore, as the case of *woman* in *The Inheritors* has shown, POS tagging cannot solve subtler problems of classification and function.

Although *has, had, are, were, will,* and *would* are all among the sixty MFW of both corpora, only the past tense forms *had, were,* and *would* appear among the sixty MFW of *Age*. And this suggests that lemmatizing the texts might help to overcome variations in tense and reduce generic differences between narration (typically in past tense), and dialogue (often in present tense). So far as I know, no large-scale tests have been performed to see how or whether POS tagging or lemmatization affects the accuracy of authorship attribution, though Burrows has long manually tagged a few very frequent words such as *to* and *that* for function before performing his analyses. As is so often true, a priori assumptions about whether either or both of these interventions would improve authorship attribution are little more than guesses. Carefully constructed tests of their effects would be valuable but extremely labour-intensive and time-consuming.

The word *I* ranks sixteenth in *Age*, suggesting a good deal of dialogue in this third-person novel and highlighting the problem of point of view. Ideally, one might compare only third person or first person texts in any one analysis, precisely because personal pronouns are so frequent and their

use varies widely in texts with different points of view (see Hoover 2001 for discussion). Unfortunately, for some problems this would drastically reduce the number of texts available for analysis, and some novels (for example, some of Conrad's) contain first person narratives within a third person frame narrative, a situation that requires tedious manual editing if points of view are to be separated. And third person novels with very large proportions of dialogue will likely diverge markedly from those with little or none. For example, *I* occurs 418 times in the first 50,000 words of *Age*, where it ranks sixteenth, just as it does in the entire novel, but only eight times in the first 50,000 words of Upton Sinclair's *The Jungle*, where it ranks 601st. Worse yet, *I* jumps from 601st to forty-sixth when the entire novel is analyzed, showing that extreme intra-textual variation is possible. Occasionally *I* ranks first in a novel, as it does in the following four eighteenth-century novels: Richardson's *Pamela* (where *the* ranks fourth, behind *I, and*, and *to*), Burney's *Evelina* (where *the* ranks third, behind *I* and *to*) and Foster's *The Coquette* and Defoe's *Moll Flanders* (where *the* ranks second). The epistolary form of the first three of these seems responsible for the extreme frequency of *I,* and *Moll Flanders* is written in the form of an autobiography (a check of six actual autobiographies finds none with *I* as the most frequent word, however).

Other personal pronouns seem potentially problematic because they are closely tied to content, and especially to the number and gender of the main characters: obviously, pronouns referring to women tend to be infrequent in texts without women. Below are the ranks of *he, his, him, she,* and *her* in several novels:

Novel	**he**	**him**	**his**	**she**	**her**
Doyle, *The Hound of the Baskervilles*	9	13	30	68	59
London, *The Call of the Wild*	4	8	14	95	75
Kipling, *The Jungle Book*	6	8	19	119	130
Doyle, *The Lost World*	12	53	14	191	238
Foster, *The Coquette*	24	26	35	15	8
Montgomery, *Anne of Green Gables*	34	85	77	11	12
James, *The Portrait of a Lady*	13	28	22	6	7
Chopin, *The Awakening*	10	24	15	6	4

In the first group, the rarity of female characters restricts the frequency of feminine pronouns. In the second, the focus on women reverses the pattern to some extent.

My own recent work attempts to cope with the difficulties mentioned above in several ultimately-related ways. I often manually remove all dialogue from novels to eliminate problems arising from differing proportions of dialogue and narration, but this requires long hours of tedious, error-prone work, and runs into difficulty in novels in which dialogue and narration are not clearly differentiated. And in some novels (*The Coquette*, for example), dialogue is not distinguished typographically, making the process very difficult and a matter of interpretation as well as analysis. Deleting the dialogue also deletes more than half of some novels, sometimes resulting in an inconveniently small sample.

I also normally remove all personal pronouns from the word frequency list before performing an analysis, or do the analysis both with and without pronouns. This is not standard practice, largely because so many analysts simply begin with the fifty most frequent words, and removing personal pronouns often gives poorer results for such analyses. Furthermore, the prevalence of male or female characters, masculine or feminine pronouns, or plural or singular pronouns may sometimes help to differentiate authors. Another innovation I have made is to delete any word that is frequent in the whole corpus because of its frequency in a single text, typically culling words for which a single text accounts for 60%-80% of all occurrences. This eliminates most proper names and other idiosyncratic items that are usually tied closely to content, though it does not eliminate main characters with the same name or place names that are frequent in two or more novels with the same setting. I

sometimes remove such words manually, but this conflicts with another innovation: using very large numbers of frequent words.

In some of my earliest work on authorship attribution and statistical stylistics, I expanded the list to the 800 MFW (Hoover 2001), and, more recently, to the 1200 MFW (http://www.nyu.edu/gsas/dept/english/dlh/TheDeltaSpreadsheets.html). Following a suggestion by Ross Clement of Westminster University (personal communication), I now typically include the 4000 MFW when working with long texts (Hoover 2005a, 2005b; Clement even reports very good results using the 6000 MFW). As we have seen for *Age,* even the sixty MFW account for almost half of the tokens of a long novel. The 1000 MFW typically account for 75%-80% of the tokens, and the 4000 MFW for more than 90%. Using such large numbers of words violates the traditional assumption that authorship markers are most likely to be found among words that are so common and ubiquitous that authors are unlikely to manipulate them consciously. It also bases the analysis on very infrequent words, a practice that statisticians are likely to frown upon. Nevertheless, dozens of unrelated analyses have shown that these large numbers of mostly content words, many of which occur only once or twice in any text, are much more effective in authorship attribution than are the 50-100 most frequent function words.

So far as I am aware, no theoretical justification exists for using such large numbers of words, but the results speak for themselves. Perhaps the main reason for the improved results is simply the much larger amount of information they provide. Whether or not the words at rank 1000 and above are individually less reliable or less informative than the fifty most frequent, there are so many of them that they both improve results and insulate the analysis from many of the errors and problems mentioned above. If *new* is left in the analysis because its mixed proper and common status goes unnoticed, there are so many other words that any small, inappropriate effect it has is overwhelmed. Similarly, if POS tagging is unavailable or not accurate enough to use, so many words with unambiguous classifications remain that the desirable but unavailable information is not finally necessary. Recent experiments suggest that using large numbers of words allows texts of different points of view, texts with varying proportions of dialogue and narration, and texts in both British and American English to be included in the same analysis without any great degradation of the accuracy of the results (Hoover 2004a). Unfortunately, large word lists are appropriate only for large texts.

Combining these methods of dealing with some of the difficulties of using word frequencies for authorship and statistical stylistics with Delta, a new measure of the difference between texts developed by John F. Burrows, seems an appropriate way of further investigating word frequencies and authorship attribution. Burrows has demonstrated the effectiveness of Delta on Restoration poetry (2001, 2002a, 2003) and has applied the technique to the interplay between translation and authorship (2002b). I have published two studies involving Delta (2004a, 2004b) that automate the process of calculating and evaluating the results of Delta in an Excel spreadsheet with macros. The first article demonstrates Delta's effectiveness on early twentieth-century novels, and shows that using the 700 or 800 most frequent words substantially improves the results achieved with smaller numbers of words, as does removing personal pronouns and culling words that are frequent in only one text. It also shows that large drops in Delta from the first to the second likeliest author are strongly associated with correct attributions. The second article shows that the accuracy of attribution can often be improved by selecting subsets of the word frequency list or changing the formula of Delta itself, as will be described briefly below. It also tests Delta and its variants on contemporary literary criticism, where they continue to perform very well. Further studies are underway by several researchers, involving a 'real life' attribution problem on nineteenth-century prose, an application of the technique and its variants to evolutionary biology, and my own projects on the style of Henry James, on narrators' styles in eighteenth- and nineteenth-century novels, and on the authorship of a late Middle English saint's life.

Consider now an authorship attribution test involving forty-six Victorian novels, by Charlotte Brontë, Collins, Dickens, Eliot, Thackeray, and Trollope. For this test, I used plain ASCII texts, downloaded from Gutenberg whenever possible, to keep the formats similar, and deleted any Gutenberg information, introductions, prefaces, tables of contents, notes, and page numbers, but not chapter or section titles, epigraphs, or dialogue. I created a word frequency list for the resulting corpus of about 5,500,000 words, selected the 7000 MFW for analysis with Burrows's Delta, and used The Delta Calculation Spreadsheet

(http://www.nyu.edu/gsas/dept/english/dlh/TheDeltaSpreadsheets.html) to calculate the percentage frequency for each word in all forty-six novels and enter a zero record for each word absent from any given text. One large novel by each author served as the primary authorial sample and the other forty novels formed the secondary test set. Before Delta was calculated, words absent from all six primary samples were removed so that means and standard deviations could be calculated. This left only 3556 words for analysis, and removed from consideration most of the character names and proper nouns of the test set.

Calculating Delta begins with the differences between the frequencies of words in the primary authorial samples and their mean frequencies in the entire primary set. To allow all of the rapidly declining frequencies to contribute equally, these differences are converted into z-scores by dividing the difference between the mean frequency of the word in the primary set and its frequency in the sample by the standard deviation of the word in the primary set. The result indicates how many standard deviations above the mean for the primary set (positive z-scores) or below it (negative z-scores) each word falls. Delta then measures the difference between test texts and primary authorial samples in a simple way. Each word's frequency in the test text is first compared with its mean frequency in the primary set, as with the primary samples. The difference between the test text and the mean is then compared with the difference between each primary authorial sample and the mean. For example, consider *the* in Charlotte Brontë's *The Professor* and the six novels that comprise the primary authorial samples:

	Bronte	Brontë	Collins	Dickens	Eliot	Thackery	Trollope
the	*Professor*	*Jane Eyre*	*Woman in White*	*Domby & Son*	*Middlemarch*	*Vanity Fair*	*Dr. Thorne*
z-score	-0.470050	-0.652670	1.398782	-0.127600	-0.861950	1.077899	-0.834460
Abs.difF.		0.182617	1.868833	0.342454	0.391903	1.547950	0.364407

In both *The Professor* and *Jane Eyre* (the primary authorial sample for Brontë), the frequency of *the* about half a standard deviation below the mean. Thus the two are about equally different from the mean, with a difference between their differences of only about .18. The differences between the frequency of *the* in the other five authorial samples and the mean are far larger, as reflected in the absolute differences. The final step is to sum the absolute differences for all words and calculate their mean, producing Delta, 'the mean of the absolute differences between the z-scores for a set of word-variables in a given text-group and the z-scores for the same set of word-variables in a target text' (Burrows 2002a, 271). The primary authorial sample with the smallest Delta is 'least unlike' it (Burrows 2003, 15) and its author is the most likely of the primary authors to be the author of the test text. Delta is extremely effective for this corpus, attributing all forty test texts to their correct authors in all analyses based on 200 or more words. Delta is so effective that it makes almost no difference whether words frequent in a single text are culled, and only in analyses based upon fewer than the 200 MFW does removing pronouns improve results. These results are so accurate that there seems little point in testing any of the alternatives to Delta on them. I turn instead to a more challenging corpus of more than 2,000,000 words of Modern American poetry, returning to poetry but moving forward to the twentieth century.

To produce results readily comparable to Burrows's, I began, as he did, with a primary authorial set of twenty-five samples by twenty-five poets and a secondary set of thirty-two samples– sixteen by members of the primary set and sixteen by other poets. I downloaded the samples from Chadwyck-Healey's Literature Online, accessed through New York University's Bobst Library, and removed section numbers, references, notes, page numbers, some dramatic sections with large numbers of character names, some dialect, all noticed foreign language passages, epigrams and other quotations, and prose sections. The resulting samples ranged from 21,000 to 72,000 words, and I took large samples for the primary set wherever possible. This resulted in a mean sample size of 44,000 words for the primary set and 38,000 for the secondary set. (One significant difference between my samples and Burrows's is that he used single long poems as the secondary texts, to more obviously mirror a real authorship attribution problem) In the best results from a preliminary test on the 20-200 MFW, those based on the 160 MFW, Delta correctly identified all samples by primary authors, but four samples by others invaded the section of correctly attributed samples. Removing pronouns improved the results slightly, but culling the word list had no effect, removing very few words overall, and none from the 200 MFW.

As noted above, much larger word frequency lists nearly always improve results, and this is dramatically true for this corpus. An analysis based on the 2000 MFW, shown in Fig. 2, is much more accurate than that based on the 160 MFW, and is the best result I was able to achieve using original Delta. Although Delta impressively attributes all sixteen of the samples by members of the primary set of authors using the 80, 100, 120 . . . 200, 400, 600 . . . 4000MFW, these results exaggerate its effectiveness somewhat. For example, as Fig. 2 shows, if the Rukeyser sample that appears surrounded by texts by others were a test text, it would not be identified correctly by this analysis, and establishing a reasonable threshold of confident attribution is difficult.

The five possible improvements on Delta proposed in *Delta Prime?* recapture information about whether a word is more or less frequent than the mean, how different the test text is from the mean, the size of the absolute difference between the test text and each primary text, and the direction of the difference between the test text and the primary text (Hoover 2004b). Delta-Lz and Delta-Oz retain Burrows's definition of Delta as the mean of the absolute differences but base the mean on a limited set of words. Delta-Lz includes only those words for which the z-score of the test text in question has a large absolute value–words much more frequent or much less frequent than the mean. Delta-Oz includes only those words for which the signs of the z-scores of the primary text and the test text being compared are opposite–words for which the test text and the primary text are different from the mean in opposite directions. Both of these measures compare each pair of texts on the basis of a newly-defined and typically unique subset of the original word frequency list. The other two measures, like Delta, include all the words, but change the definition of Delta itself. Rather than taking the mean of the absolute differences between the test text and each primary sample, Delta-2X doubles the mean of the differences that are positive (words for which the z-score of the test text is greater than that of the primary sample–words more frequent in the test text than the primary sample) and subtracts the mean of the negative differences (words less frequent in the test text than the primary sample). Because the second figure is a negative, this calculation in effect adds the absolute value of the negative mean. Delta-3X triples the positive mean before subtracting the negative mean. Finally, Delta-P1 adds one to the positive mean and squares that sum before subtracting the negative mean. All three of these measures weigh more heavily those words that are more frequent in the test text than the primary sample, treating presence as more significant than absence.

Delta-2X improves slightly upon Delta, and Delta-3X and Delta-P1 do better still, though in all cases one member's text falls below the most obvious threshold and would not be correctly identified if it were being tested. Both Delta-Oz and Delta-Lz (>0.7) produce results (shown in Figures 3 and 4) in which none of the samples by others invade the samples by members, but only Delta-Lz shows a clear threshold, and two members fall below it. Delta and the Delta Primes clearly reflect real authorship information and good results can be replicated on diverse sets of texts, but correctly attributing texts by members is not enough–the analysis must avoid falsely attributing texts by others to members of the primary set. That no samples by others invade the samples by members is encouraging, but this kind of testing seems inadequate. Over-interpretation of results is all too easy when the truth is known in advance.

Beginning with a large group of samples of poetry, many by the authors tested above and some by additional authors, and selecting smaller samples to increase the difficulty in the hope of revealing any differences in effectiveness among Delta and the Delta Primes, I created a simulation in which the true authors were not initially known. I first selected twenty-five primary samples, eleven by members and thirteen by others. I then added fifteen more samples (six by members and nine by others), hid their identities, put them in random order, and had a helper who knew nothing about the simulation select twelve of the fifteen to rename TEXT01 . . . TEXT12, by authors A01 . . . A12. For the simulation, the mean sample size was about 24,500 words for primary samples, 22,000 for samples by members, 13,500 for samples by others, and 16,000 for the test samples. The simulation left me with the unwanted knowledge that at least three and no more than six samples were by members, but the results in Figures 5-10 show that this knowledge could hardly have affected the results. A precise characterization of the relative effectiveness of Delta and the Primes is difficult, but all of the Primes, with the possible exception of Delta-P1, seem more effective than Delta.

These results do not show the whole picture, however. Analyses involving the 200, 400 . . . 4000 words for Delta and the five Primes, 120 analyses in all, produced only twenty-one errors, never more than one in any analysis. Fourteen of the errors involved Delta-P1, which also produced the

only analyses in which test samples other than 1,3,4,8, and 11 mingled with the samples by members. Finally, the following five attributions were absolutely consistent across all 120 analyses: TEXT01 = Frost, TEXT03 = Gunn, TEXT04 = Sandburg, TEXT08 = Creeley, and TEXT11 = Kooser. None of the suggested attributions of texts known to be by others attained this kind of consistency, though Rich was frequently the most likely author for TEXT12, as Hart Crane was for TEXT09. (One or two authors frequently appear as the most likely author for many of the texts by others in an analysis.) Finally, as Fig. 11 shows, the changes in Delta and Delta-z from the likeliest to the second likeliest author were generally much greater for these five test texts and the samples by members than for the other seven test texts and the texts by other authors (the texts are arranged in order of increasing change in Delta).

At this point it will come as no surprise that these five attributions are correct. As Figures 5-11 show, however, if the Rich sample that appears as a member text had instead been a test sample, none of the analyses would very convincingly have attributed it to her. Only Delta-Oz and Delta-Lz, however, strongly suggest that this sample is not by a member of the primary set. Perhaps these two Primes are more sensitive to intra-author variation than are Delta and the other Primes, but that is a topic for further research. It would be tempting to suggest the small size of this sample compared to Rich's other samples as a cause (it is only about one third as long), but the two strongest attributions in all of the analyses shown above are the two other pairs with the greatest discrepancy in size. A more promising explanation is suggested by the fact that the larger samples come from *Collected Early Poems: 1950-1970*, while the smaller one comes from *Midnight Salvage: Poems, 1995-1998*, and Rich notoriously changed her style during her long career. This 'failure' reminds us that some authors' styles change dramatically over their careers, and that some authors use very different styles in different texts. This calls into question the basic assumption of an invariable wordprint for each author and suggests that statistical methods alone may not always be adequate and may need to be augmented by more time-consuming methods such as those based on extreme differentials in the use of pairs of word.

Delta and the Delta Primes are extremely effective in attributing the five samples by members and in rejecting the seven samples by others. The fact that the failures are all failures to attribute rather than failures of attribution is important: in no case does Delta or any of the Delta Primes strongly encourage a false attribution. Further tests on contemporary prose and on samples that are lemmatized or tagged for part of speech would be helpful–not so much to confirm the effectiveness and reliability of Delta and the Delta Primes, which now seem very solidly validated, but rather in the hope of more fully understanding why these relatively simple techniques work so well, and in continuing to improve their already impressive power.

References

Burrows, John F. , 'Questions of Authorship: Attribution and Beyond', paper given at the ACH/ALLC, Joint International Conference, New York, June 14, 2001.

Burrows, John F., '"Delta": a Measure of Stylistic Difference and a Guide to Likely Authorship'. *Literary and Linguistic Computing,* 17 (2002), 267–287. (2002a)

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

Burrows, John F., 'Questions of Authorship: Attribution and Beyond', *Computers and the Humanities* 37:1 (2003), 5–32.

Hoover, David L., 'Statistical Stylistics and Authorship Attribution: An Empirical Investigation'. *Literary and Linguistic Computing* 16:4 (2001), 421–444.

Hoover, David L., 'Testing Burrows's Delta', *Literary and Linguistic Computing,* 19:4 (2004), 453–475. (2004a)

Hoover, David L. ‘Delta Prime?’ *Literary and Linguistic Computing,* 19:4 (2004), 477–495. (2004b)

Hoover, David L. ‘Delta, Delta Prime, and Modern American Poetry: Authorship Attribution Theory and Method’, paper given at the ALLC/ACH Joint International Conference, University of Victoria, Victoria, BC, Canada, June 16, 2005. (2005a).

Hoover, David L. ‘The Delta Spreadsheet’, paper given at the ALLC/ACH Joint International Conference, University of Victoria, Victoria, BC, Canada, June 17, 2005. (2005b).