A Second Opinion on "Shakespeare and Authorship Studies in the Twenty-First Century"[1]

John Burrows

POLONIUS What do you read my Lord?
HAMLET Words, words, words . . . [2]

I. Introduction

One of Brian Vickers's energetic review essays is directed at computational stylistics as a scholarly enterprise and as exemplified by a recent book. [3] It amounts, however, to an exercise in self-exposure. It seems rather soon to speak so generally of any aspect of our new century. Even within Shakespeare studies, few would accept his appraisal of emerging methods of attribution as a winner-takes-all combat between the statistical analysis of relative word frequencies and Vickers's method of gathering collocations. He attacks the former and promotes the latter. The attack fails because he does not understand the methods he derides or even pay attention to things plainly stated in the book under review and in other work that he cites. My own "second opinion" aspires (as in medical practice) to be better informed on the matters at issue but not to encompass the whole field of recent authorship studies.[4]

I am indebted to Hugh Craig for access to his collection of electronic texts and his software package, "The Intelligent Archive." Since the tests undertaken here were of my design and the graphs of my making, I take full responsibility for any errors, oversights, or blemishes. (Sir Brian Vickers makes the point that anyone who uses his texts and his software can replicate his results. That is a virtue he shares with Craig, whose material is available on request.) I am also indebted, for helpful suggestions, to two friends and to the readers chosen by the editor of this journal.

[1] See Brian Vickers's review essay, "Shakespeare and Authorship Studies in the Twenty-First Century," *Shakespeare Quarterly* 62 (2011): 106–42.

[2] *Hamlet*, 2.2.190–91. Quotations are from *The Complete Works of William Shakespeare*, ed. Herbert Farjeon, intro. Ivor Brown, 4 vols. (London: Nonesuch Press, 1953). Act, scene, and line numbers are from *William Shakespeare: The Complete Works*, ed. Peter Alexander (London: Collins, 1951).

[3] Hugh Craig and Arthur F. Kinney, eds., *Shakespeare, Computers, and the Mystery of Authorship* (Cambridge: Cambridge UP, 2009).

[4] There have been several worthwhile surveys of the field in the last twenty years. So far as I am aware the most recent and most wide-ranging assessment is that of Efstathios Stamatatos,

I begin with two disclaimers. My long and friendly association with Hugh Craig prevents me from offering an impartial assessment of his work. Nor can I claim to be a Shakespeare scholar. As reader and playgoer, student and teacher, I have been enriched by my acquaintance with Shakespeare and his fellows for almost seventy years, but unlike Craig, Kinney, and Vickers, I have not given the greater part of my academic life to the study of Renaissance drama. In a third matter, however, I am on firmer ground. Much of Vickers's attack is upon methods of analysis with whose development I have been associated. Here, again, I cannot claim to be impartial. But, since Vickers is so kind as to call me the "progenitor"[5] of these methods, I feel free to defend them and to offer him an equally amiable gesture in return.[6] An opening sketch of the scene may be of service.

II. Questions of Method

Words and Numbers

It is among life's ironies that, throughout the international community of humanities scholars, so many of us are uncomfortable with numbers. For, at the heart of our customary work are tenets that we share with statisticians. Like us, they deal in the comparative and not the absolute. We, too, regard the balance of probability as an invaluable guide. We do not often expect our questions to yield finality but rather to open further questions. We do not expect any one form of evidence or any one line of reasoning to suffice and, accordingly, we look for corroboration. We differ markedly from statisticians in not attempting exact calculations of probability. Where we offer the iteration of examples as evidence, the statistician can transform a vast array of data into succinct and lucid graphs. But we, too, seek signs of resemblance and difference, presence and absence, concurrence and divergence. As a traditional literary scholar who turned to numerical evidence in the latter part of a long career, I have found new and helpful ways of approaching old questions and opening up some new ones. Many interesting things cannot be counted but many others can. The remarks offered here are intended for readers unfamiliar with the analytical methods at issue. Much of what is said now will be amplified later in the discussion, and proper references will be offered for those who care to pursue them. But these matters are less arcane than many people suppose.

"A Survey of Modern Authorship Attribution Methods," *Journal of the American Society for Information Science and Technology* 60 (2009): 538–56; online at http://www.icsd.aegean.gr/lecturers/stamatatos/papers/survey.pdf (accessed 16 February 2012).

[5] Vickers, "Authorship Studies," 119.

[6] In order to rebut Vickers's misrepresentations of it, I must refer more often than I would wish to my own published work. This does not mean that I regard myself as single in the field, a Solitary Old Reaper.

Most statistical analysis rests upon straightforward principles of comparison and on simple arithmetic. (The technical complexities usually have to do with ensuring that comparisons are as equitable as possible.) Whether it is tabulated in sets of scores or represented in graphs, the material treats of *specimens* that are compared on the basis of appropriate *variables*. In authorship studies, the specimens are texts or parts of texts appropriate to the case in question. The variables can comprise any countable form of textual phenomena that may admit authorial differences. Sentence length, word length, relative frequencies of different words across a range of texts, vocabulary richness, and other more esoteric phenomena have all been studied over the last century and a half. The collocations counted up by Brian Vickers are a modern variant of the parallel passages studied by scholars down the centuries. Advances in computing over the last forty years or so have vastly increased the scope, accuracy, and speed with which all these phenomena can be investigated. These changes, however, need not and should not compromise the conventions of rational inquiry.

The simple logical principle of *presence-and-absence* yields one useful approach to the comparison of texts. Since classical times, keen-eyed scholars have identified *hapax legomena*, words that occur once only in a disputed text, and have regarded their presence in other texts as a sign of shared authorship. We are now able to compare large bodies of texts and identify all the *hapax legomena* that distinguish one set from the other—Shakespeare, let us say, from Jonson (and vice versa) or even Shakespeare from non-Shakespeare. The proportion of *hapax legomena* in a text diminishes slowly as its length increases. At 10,000 words, they may number 800 or so, whereas 50,000 may yield only 2,500—a slide from 8 to 5%. We can readily extend the range to embrace all such words, whether frequent or once only, as occur only in each of our main authorial sets. Working from these two lists of words, we turn to fresh specimens of each author's work and finally to the text whose authorship is disputed. To which group does it better conform? From which does it more markedly diverge? As may be supposed, many of the words chosen in this way do not recur in fresh texts by the author who used them. But far more of them do so than appear in fresh texts by the author who previously did not use them. It is obviously desirable to prevent words associated with a given topic (and not really characteristic of an author) from playing too strong a part. We find it best, therefore, to register each word's presence in a given specimen as a single "hit" and to take no account of its further occurrences there. On each of the two lists, the total number of such hits per thousand words represents the outcome for each specimen. While this test, known as Iota, figures only briefly in what follows, it illustrates a useful form of reasoning.

Even among the comparatively uncommon words, *consistency of use* offers a second mark of authorship. If we separate each of the aforesaid large sets of texts into five successive parts, we can select such words as occur in three, four, or all five parts of the one set and in only one or two parts of the other. (Those that never occur in the other can be included but are better kept aside for corroborative use in the Iota test.) In my original version of this Zeta test, I offered separate results for different levels of consistency. Hugh Craig's Zeta variant, an improved version much used in Craig and Kinney's book, registers simple *ratios* of consistency and picks out those words that offer the strongest ratios. After the two word lists are established, the procedure is as described for Iota. Both Zeta and Iota can be used to test one authorial candidate against many but are at their best in head to head contests between two candidates.

Two other tests, both better known than these recent offerings of my own, deal in *relative differences of frequency* among words freely used by everyone who writes in English (and likewise among the peers of such words in all the languages in which these tests have been applied). Both of these tests are especially useful when many authors are to be compared. The Delta test assesses levels of divergence from a common base by taking each text specimen in turn and *averaging* its divergences on a large range of variables. Which specimens differ least from a disputed text? Principal component analysis (PCA) assesses *concomitant variations of frequency* across a range of texts. This process allows the texts to display their various affinities and disaffinities, enabling us to form (and subsequently verify or falsify) inferences about the manner in which they array themselves.

No more need be said of Delta at this point. It is not used by Craig and Kinney, and Vickers dismisses it. But PCA is a staple of Craig and Kinney's argument, and it plays a large part in Vickers's attack on computational stylistics. It is a statistical test of long standing, introduced by Karl Pearson in 1901 and since used for different purposes in many different fields. In order to appreciate its place in studies of attribution, let us briefly consider the frequency hierarchy of the English vocabulary and the operation of concomitant variation.

As is well known, *the* is the most frequent of all English words. In Hugh Craig's archive of 202 Renaissance plays, this single word type makes up 3.01% of all the word tokens used.[7] Three in every 100 words, 600 in a play of 20,000 words. And yet that is not so many. In a more diverse range of texts, *the* is likely to make up about 5% of all words; in descriptive prose, fictional or otherwise,

[7] Students of linguistics call the various instances of a given word-form (such as *the* or *and*) "word-tokens," examples of a given "word type." In *the cow jumped over the moon*, we have six tokens but only five types.

it can run as high as 9%. Its relative infrequency in drama reflects one of many differences between dialogue and description. By the same token, the dramatists make much more frequent use of *I* than do novelists like Scott, Mary Shelley, and Wilkie Collins. At 2.73%, *I* ranks third (after *the* and *and*) in the word list derived from Craig's drama archive. In passages of impersonal narrative, on the other hand, it may not occur at all.

Even in these two word variables, the principle of concomitant variation begins to manifest itself, with scores for *the* and *I* pursuing opposing but roughly symmetrical paths across most members of a given range of (let us still say) plays and novels. When such scores are subjected to the statistical procedure known as correlation, a high negative correlation coefficient will emerge. These coefficients run from -1.00 to 1.00, with the extremes representing perfect correlations rarely seen in the real world.[8]

But *the* and *I* do not stand alone. When further word variables are introduced, *the* finds allies in most prepositions, which yield strong positive correlations with it and with each other. *I*, meanwhile, correlates very strongly with *my* and *me* and often with *you* and *your*. The relative pronouns, except sometimes for *that*, tend to correlate positively with *the* and prepositions but negatively with dialogic pronouns. It need hardly be said that most nouns, adjectives, and adverbs lean rather to the descriptive and "thing-ish" than to the dialogic. On the whole, the more frequent conjunctions behave less predictably.

It should be noted at this point that concomitant variation is not a property peculiar to the more frequent word variables. In Craig's word list, the average of 3.01% for *the* stands well above 1.92% for *of*, 0.31% for *at*, and 0.18% for *upon*. In a play of 20,000 words, these yield averages of about 600, 380, 62, and 36, respectively. Disparate as they are, these scores can all yield positive correlations with each other because they can all be expected to show roughly similar rises and falls across our set of texts. While it is true that such patterns become more ragged in the lower reaches of a word list, they rarely disappear.

[8] As an illustration, let us take the case of a four-way intersection between two streets with traffic lights at each point of entry. The traffic can flow north, south, east, or west (but for simplicity we shall not allow vehicles to turn from one street into the other). In a world where there were no breakdowns, collisions, or bad drivers, the traffic lights would ensure that, when vehicles traveling north-south or south-north were in motion, all those traveling east-west or west-east would be stationary, and vice versa when the lights changed. All that amounts to concomitant variation across a range of specimens, large or small, fast or slow. It would yield a positive correlation of 1.00 among the vehicles travelling either way in either street, and a negative correlation of -1.00 with those of the other street. But human behavior does not admit such mathematical perfection. And although the English language is systematic, it is certainly not mechanical in its regularities. We find that 0.8 or thereabouts is as close a correlation as can reasonably be expected in any ordinary set of word frequencies.

As the number of word variables being tested is increased by moving further down the word list, the simple contrast of genres between plays and novels is reinforced by changes in the language from Elizabethan to later times. *Thou / thy / thee*, *art / hast / dost*, and *hath / doth* all occur much more freely in Elizabethan plays than in Victorian novels and help to strengthen the concomitant variations among the different texts.

When two groups of texts differ so obviously from each other, strong contrasts are only to be expected. Let us set the plays aside and focus on the novels of three centuries. They should still differ among themselves by virtue of changes in the language over time and by virtue of marked differences in their relative proportions of dialogue and narrative. And so they do. Further contrasts among the novels mark differences in the authors' education. Those authors, mostly male, who had a classical education write more formally and use a more Latinate vocabulary than those authors, mostly female, who did not. As we come closer to our own times, the language of novels in English also differs according to the nationality of the authors. All these differentiae manifest themselves in concomitant variations of word frequency between different subsets of specimens.

In the simple case with which we began this part of the discussion, it is easy to see that *the / I* yields a correlation coefficient of perhaps -0.75, *the / of* +0.70, and *I / my* of +0.80. Easy, too, to consider why that should be. But a matrix (or table) of coefficients in which a hundred word variables are correlated, each with every other, offers more information than our minds can register. It is still easy enough to locate examples of known interest or to pick out a handful of the strongest coefficients, positive or negative. But the matrix also contains a huge and bewildering array of other interrelationships, some strong and meaningful, others too weak to matter. At this point, we can turn to PCA for help. The procedure seeks *patterns* in a correlation matrix and identifies them as a succession of principal components (PCs) in diminishing order of their strength. In each successive component, all the variables are arranged in a ranked series or vector. In many applications of the procedure, such as engineering, the differentiae governing each vector are known in advance and the question at issue bears on the behavior of specimens. What is the maximum velocity attainable beneath a safe temperature ceiling? What is the least thickness needed to bear a given load? In its application to the evidence of word variables, however, the differentiae governing the successive vectors are not known in advance, and we must form inferences about them by observing their patterns and drawing upon our understanding of the language and our knowledge of the current set of text specimens. The specimens can then be matched against the variable pattern, usually in the form of a scatter plot showing how they arrange themselves on the basis of the first two PCs, the two most powerful vectors.

In the case where Elizabethan plays and nineteenth-century novels were compared as whole texts, the first vector would yield two separate clusters of specimens distinguished by the immense difference of genre. Within each cluster, however, some members would lie nearer the borderline than most. These would include novels, like Jane Austen's, in which the proportion of dialogue is high, and novels, like *Wuthering Heights*, where much or all of the narrative itself is couched in the first person. In the other cluster, some plays, like Shakespeare's tragedies, would show the effect of a more descriptive and reflective dialogue than most. Meanwhile the second PC would yield a rough chronological clustering, reflecting changes in the language over the centuries.

If the novels were omitted and the plays were taken alone, differences between tragedy and comedy would yield the strongest separation, with *thou*/*you* and their affiliates strong in the brisk dialogue of most comedy while the more reflective style of much tragedy would be marked by words of a connective function. Historical changes in the language from 1576 to 1642 would manifest themselves in the second vector.

If, on the other hand, the plays were omitted and the novels were taken alone, the contrasts between the language of dialogue and that of narrative would continue as one major differentia, as would the effect of change over time. But a corpus of fiction by a hundred authors born from 1660 to 1954 embodies other differentiae than these. In this corpus, we have almost as many female authors as males; and, from the late eighteenth century onward, regional and national differences also emerge. So far, we have allowed our word list to rule the game without intervention, thus enabling the text specimens to arrange themselves according to their various affinities and disaffinities. Is it possible, we may now ask, to identify patterns of word frequency that separate male novelists from female or Britishers from Americans and Australians?

Such attempts require us to turn from the exploratory function of PCA to its possible use in exercises of *classification*. To test the contrast between such subgroups, it is obviously desirable to focus the inquiry on either narrative or on dialogue, thus excluding one major differentia. For my own work in this area,[9] I introduced a second refinement, confining my investigations to histories (as first-person retrospective narratives are often called) and to authors whose work yields at least three texts of that kind. For the study of contrasts between the different nationalities, those British authors whose work appeared before that of the earliest American and Australian writers were excluded. At this stage,

[9] J. F. Burrows, "Tiptoeing into the Infinite: Testing for Evidence of National Differences in the Language of English Narrative," in *Research in Humanities Computing 4*, ed. Susan Hockey and Nancy Ide (Oxford: Clarendon Press, 1996), 1–33.

incipient national clusters were apparent in the analyses but the picture was still clouded.

Ever since he introduced it in 1908, Student's *t*-test, so named by its inventor William Sealy Gosset, has been a standard method for comparing the behavior, on a range of variables, of two putative subgroups of specimens. Working on each variable in turn, it compares the mean scores for each subgroup, takes account of their standard deviations,[10] and yields a numerical expression of the probability that the two subgroups are *not* separated by that variable. Where probability (p) = 0.1 or more, there is 1 chance in 10 that the groups are not separable, and it is usual to dismiss that variable as of no evidential weight for that purpose. At p = 0.05, or 1 in 20, statisticians begin to speak of a significant result; their interest increases rapidly as the odds extend to 0.01, 0.001, and beyond. The number of variables studied must also be taken into account. If there are a hundred variables, more than five variables where p = 0.05 and more than one variable where p = 0.01 would be required. If no more than that occurs, the subgroups are still not to be regarded as separable on that evidence. And if the evidence is to carry any real weight, these minima need to be surpassed by a large margin. (Something over twice the minimum allows for a technicality called the two-tailed effect and also for a margin of error.)

In one study of the first-person narratives described above and testing the hundred most frequent variables (such as *the*, *and*, *I*, *of*, and *a*), these requirements were easily satisfied.[11] Twenty British authors born before 1780 were compared to twenty-three born after that date. Of the 100 variables, no fewer than 57 passed the 1-in-20 test of p = 0.05; 42 were rated $p \geq 0.01$. For corroboration, I split the same forty-three authors into two groups by the alphabetical order of their surnames. Beginning again with the same alphabetical sequence, I counted off names as either odd or even and grouped them accordingly. Not surprisingly, the results showed no significant separation between these arbitrary subgroups. Only 5 of the 100 variables in the one case and only 2 in the other

[10] Standard deviations are a measure of relative divergence from a mean score. They increase in size when the mean rests upon widely scattered scores. Speeds on a German *autobahn* or in a mass sprint race would diverge much more widely from the overall mean than would speeds in busy urban traffic or in a championship sprint. In the former cases, moreover, distinct "populations" might be distinguishable: high-performance cars and old bangers; champion runners and strugglers. The more volatile the scores, the higher the standard deviation in proportion to the mean. In a hypothetically perfect distribution, around 68% of all scores would lie between 1.0 and -1.0, that is, within a range a single unit of standard deviation on either side of the mean. 95% would lie within two units either way and 99.7% within three.

[11] For the results discussed in this and the following paragraph, see Burrows, "Tiptoeing," 16–21, Tables 2.1–2.3, 3.1–3.3.

exceeded values of $p = 0.05$. In this investigation, at least, chronological change in the language exerts powerful effects; a *reductio* yields absurdity.

In my comparisons of authors from different countries, it was desirable to match contemporaries. For this reason, the British authors born before 1780 were excluded. Comparisons between twenty-three later British authors and twenty Australians, between the same Australians and twenty U. S. authors, and then between the British and the American authors yielded results in which, in each case, the 100 variables yielded around 20 variables at rates better than $p = 0.05$. The national differentiae were less powerful than the chronological but were still of real weight.

When PCA is applied to these various pairs of subgroups and the original long word list is replaced by the select set of marker words (picked out by the *t*-test), the clusters are more clearly separated from each other and provide a better basis for testing fresh specimens of any given provenance. PCA has thus been "optimized for classification," a phrase to which we shall return. To put it another way, classificatory exercises like these benefit from a reduction in the "noise" that obscures the "signal." For information scientists, noise is defined by the needs of the task at hand: the noise that obscures one message may well be a message in itself. Some of the word variables that best distinguish American authors from the British do nothing to distinguish the Australians from either of those parties. When these variables are set aside, it is not as valueless but is not relevant at present.

In the early 1990s, the potentiality of such work led me to take up the term "computational stylistics" to distinguish it from the head-to-head comparisons that prevailed in what was called "stylometry." While the older term is still widely used, the distinction I had in mind has been secured by many subsequent advances.

But what of attribution studies? Since work on attribution now admits comparisons among many candidates and can be based on the evidence of word frequencies, the tests I have sketched in this section can all be brought to bear on questions of this kind. A list of word variables is compiled, tried out, and then sifted to fit the case in hand. The body of text specimens used for this purpose provides "training sets" against which the behavior of the disputed text can be assessed. Given a sizable body of texts suitable for comparison, worthwhile evidence can usually be elicited. The task is at its most difficult with authors like John Dryden, whose stylistic repertoire exhibits unusual versatility; Edmund Waller, whose work changes radically over a long career; and an occasional hard case like Thomas D'Urfey, whose style is almost too colorless to yield a legible "signature." Unexpectedly perhaps, the stylistic repertoire of the man in the street is usually so narrow as to be easily distinguished from other styles.

By way of support for the claims made in this section, Craig and Kinney's book *Shakespeare, Computers, and the Mystery of Authorship* offers many successful examples; below, I add some of my own. In the meantime, we must attend to the case that Vickers puts.

The Present Case

Vickers opens his review essay with a straightforward and not unfavorable account of a lately published (but actually much earlier) inquiry into a celebrated attribution problem. He ends with a glowing account of an approach that he is using in current work of his own. This method deals in "*N*-grams," little collocations of several words as markers of authorship. While his variant is not without promise, it is less innovative than he suggests.[12] As with our methods, it should benefit from refinement through experience. As I shall show, it needs more disciplined management in order to ensure that no authorial candidate is given favored treatment. But this way of treating *N*-grams may well become a useful new instrument to complement those we already have. It appears, however, that this beginning and this ending are partly designed to serve as rhetorical flourishes—specimens of the acceptable and the truly admirable, respectively—against which the alleged failure of computational stylistics can be seen in high relief. Let us turn, then, to the long middle section of Vickers's review essay.

We are told that Craig and Kinney cannot see grave weaknesses in their methods and a crippling unreliability in their results. As the alleged progenitor of such error, I too am much at fault. We are a crew of rude mechanicals who do not read our texts. Our work is not supported by any theory of language. And, blinded by our absurd belief that individual words are discrete entities, we do not even see that they cannot form the fabric of discourse without being interwoven. Such charges must be answered.

Evidence Misread

Let me begin with two examples to justify my earlier claim that Vickers does not give due attention to things plainly stated. They have to do, respectively, with two of Craig and Kinney's figures (the first of which is reproduced here as Figure 1),[13] their adjoining commentary, and Vickers's response. Both examples are important because they assess the efficacy of one of the methods that are put to use in cases where the truth is doubtful or unknown. Vickers describes the procedure itself as "one of Burrows's more recent (2007) methods, called 'Zeta.'"[14]

[12] Stamatatos, "Survey," section 2.1 (pages 3–6).

[13] Hugh Craig and Arthur F. Kinney, "Methods," in *Shakespeare, Computers*, 15–39, esp. 22, 24 (Figures 2.2 and 2.3).

[14] Vickers, "Authorship Studies," 119.

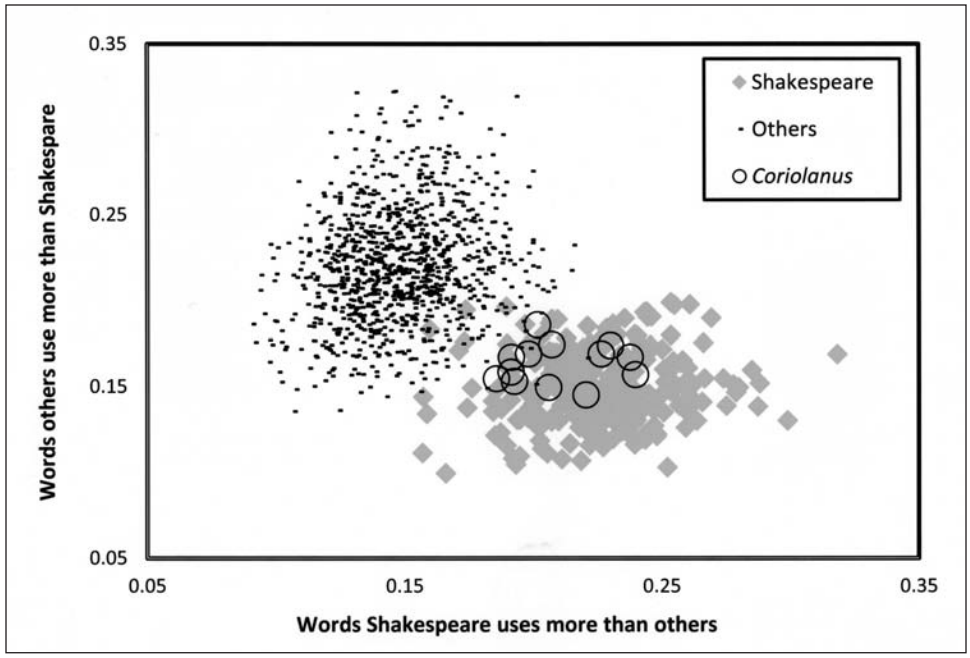

Figure 1: Shakespeare, other dramatists, and *Coriolanus*. Reproduced with permission from Hugh Craig and Arthur F. Kinney, "Methods," in *Shakespeare, Computers, and the Mystery of Authorship* (Cambridge: Cambridge University Press, 2009), Figure 2.2, page 22.

Strictly speaking, the test used here is the improved version, Craig's Zeta variant, as sketched above (see page 358).[15]

Figure 1 is a scatter plot embracing three groups of entries: a cluster of grey diamonds representing 278 segments of 2,000 words apiece from Shakespeare's plays; a cluster of black dots representing 1,009 such segments from the plays of his contemporaries; and a little cluster of thirteen circles representing just such segments of *Coriolanus*, a play chosen at random and then kept aside for testing against the larger groups.

The mainspring of Vickers's assessment is his claim that "seventeen of the *Coriolanus* segments fell on the Shakespeare side of an imaginary straight line between the two clusters, but five (nearly 30 percent) did not."[16] If this were true, Vickers's scorn would be well founded. But it is not. His 5 out of 17 (not 22) is an unfortunate miscalculation, but it sets the tone for what follows. When

[15] The original form is described in my 2007 article, "All the Way Through," to which Vickers refers; see "Authorship Studies," 115n27. The variant is described in Hugh Craig and Arthur F. Kinney, "Glossary," in *Shakespeare, Computers*, 223–27, esp. 226–27.

[16] Vickers, "Authorship Studies," 119–20.

his claim is put aside, the 5 and the 17 still have nothing whatever to do with *Coriolanus*. The 5 (out of 278) are aberrant Shakespeare segments that stand in non-Shakespeare territory. The 17 (out of 1,009) are the non-Shakespeare segments that stand in Shakespeare territory. So Craig and Kinney rightly claim a 98% rate of success in separating the two clusters of their "training set." When they do turn to *Coriolanus*, the test set, in their next paragraph, they note that the circles, which represent the 13 segments of that play, all fall in Shakespeare's territory. Not unnaturally, they regard this outcome as a successful trial of the efficacy of their test.

A second figure from Craig and Kinney's study, not reproduced here, is a scatter plot formed in the same way.[17] It is designed to test a randomly chosen non-Shakespeare play, Middleton's *Hengist, King of Kent*. The aggregate number of entries in the two main clusters is altered slightly by the removal of *Hengist* from one set and the addition of *Coriolanus* to the other. This time (on figures supplied at my request by Craig), there are only 4 aberrant Shakespeare entries and 17 non-Shakespeare—as before, a 98% success in separating the two clusters of the training set. This time, however, 1 of the 10 entries for the test play strays into the wrong territory. Altogether, that makes 22 correct answers out of 23 for *Coriolanus* and *Hengist*—a trifle above the 95% level of accuracy to which we are accustomed.

What draws Vickers's notice here is a pair of striking aberrations in the training set. One member of each main cluster lies deep in the opposing territory. Sweeping aside what Craig and Kinney have to say, Vickers comments, "The contributors to this volume place too much reliance on methods which, in the current state of knowledge, are not fit for purpose."[18] But let us seek a brief stay of execution while we consider the force of probability. All told, in training set and test piece, we have a total of 22 errors, 2 of them egregious. Over 1,300 entries are correctly placed. A proper assessment must give due attention to both sides of the balance. Few of us, in any facet of our lives, would reject odds that favored us by 1,300 to 2 (or even 22). While knowing that we might lose, we should rightly expect to win.

But Vickers persists, pressing what sounds like a reasonable objection: while this may be acceptable in cases where the truth is known, attribution deals in cases where it is not. In such cases, we have no way of telling how many entries are erroneous or any way of identifying them. But errors in the training set are not open to this objection. We know how many there are, we can identify each one of them, and we are free to decide in advance whether the separation of the

[17] Craig and Kinney, "Methods," Figure 2.3.
[18] Vickers, "Authorship Studies," 121.

main clusters suffices for our purpose. The erroneous entry for *Hengist* is what matters. If there were any doubt about Middleton's authorship, the notion that the stray segment was actually by Shakespeare would need to be entertained. That is why corroborative work with independent tests is essential. Even then, since we are dealing with statistics, the same result might recur: lightning can strike in the same place, but by now the odds in favor of the truth would be immense. Yet Vickers shows no interest in corroboration: "Nothing new there," he says wearily at such times.

A Comedy of Errors

I turn now to larger questions of method. By way of giving bottom to his dismissal of computational stylistics, Vickers adopts Joseph Rudman's notion that "nontraditional" attribution methods have failed because thirty years of work have not yielded a "killer app"—a single approach accepted by everybody and applicable in every case.[19] No such approach has yet emerged, I respond, because internal evidence, traditional or otherwise, cannot be expected to yield one. The introduction of statistical analysis into literary studies merely highlights an ancient truth: with internal evidence we form our conclusions on the balance of probability. And that balance tilts in our favor when several independent tests yield mutually corroborative outcomes at high levels of confidence. That is why we must keep trying out the tests in cases where the truth is known. Seen in this light, the work of the last thirty years is far more fruitful. In the 1980s we strove, even under favorable conditions, to reach 80% levels of accuracy. We now have several independent tests that regularly surpass 95% on texts of 2,000 words or more. Evidence of this can be seen in Craig and Kinney's book; in articles of my own, including those that Vickers cites; and elsewhere in the scholarly periodi-

[19] Rudman has given many conference papers in this vein. For an abstract of the latest, see Joseph Rudman, "The State of Non-Traditional Authorship Attribution Studies 2010: Some Problems and Solutions," http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/pdf/ab-596.pdf (accessed 30 November 2011). In both his conclusion and his list of references, Rudman alludes to an article of which I was joint author and to his own "riposte". He says nothing of my full and firm reply to him in the next issue of the same journal. Coming as it does from our self-proclaimed gatekeeper of scientific standards, this omission speaks eloquently for itself. For those so minded, the document trail is as follows: John Burrows, "Sarah and Henry Fielding and the Authorship of *The History of Ophelia*: A Computational Analysis," *Script and Print* 30 (2006): 69–92; Anthony J. Hassall, "Sarah and Henry Fielding and the Authorship of *The History of Ophelia*: Literary Considerations" *Script and Print* 30 (2006): 93–100; Joseph Rudman, "Sarah and Henry Fielding and the Authorship of *The History of Ophelia*: A Riposte," *Script and Print* 31 (2007): 147–63; and John Burrows and Anthony Hassall, "Sarah and Henry Fielding and the Authorship of *The History of Ophelia*: A Reply," *Script and Print* 31 (2007): 220–29. The journal editor's comment at the end of my reply indicates that Rudman had seen it and chosen to say no more (229).

cals, especially *Literary and Linguistic Computing*. The files of that journal show a surge of such activity in the last decade. Vickers says nothing of all this, not even offering a gesture to new work on Shakespeare by his long-standing allies, Ward E. Y. Elliott and Robert J. Valenza.[20] And the progress is continuing. At a July 2011 symposium held in Newcastle, Australia, Tomoji Tabata of Osaka University presented a study of Dickens in which a new method called "Random Forests" yielded 98% accuracy and offered a clear display of the evidence on which it differentiated among the specimens.[21] If Vickers's use of *N*-grams proves as successful, it will take its place in good, serviceable company where restless phantasms of "killer apps" never trouble the mind's eye.

Vickers also draws on Patrick Juola's "damaging criticisms"[22] of PCA. He does not mention that it is among the methods used by Juola and his colleagues.[23] As experienced practitioners, they would know that every method has strengths and limitations; that the limitations of one can be offset by the strength of another; that, *pace* Vickers, PCA can be "optimized for the task of classification"[24] by using Student's *t*-test to assist it; and that statistical results are always inferential. Knowing also that PCA does not lend itself to the calculation of error rates, they would appreciate that the actual errors in a given trial are plainly visible in the scatter plots. On some related matters, Vickers plays naïve (or *faux-naif*) semantic games: "Craig and Kinney rightly describe PCA as a data reduction method: it cannot also be a suitable method of classification."[25] But it should be noted that data reduction is a method of "noise suppression." It picks out the strongest patterns so that they may be assessed as salient or not. Vickers accuses Craig and Kinney of "even failing to register the caveat of their scholarly progenitor, John Burrows, who explicitly warned that 'PCA is not intrinsically a test of authorship but only of comparative resemblance.'"[26] Certainly, as Vickers affirms, resemblance is not identity; that is why Craig and

[20] Ward E. Y. Elliott and Robert J. Valenza, "Two Tough Nuts to Crack: Did Shakespeare Write the 'Shakespeare' Portions of *Sir Thomas More* and *Edward III*?" *Literary and Linguistic Computing* 25 (2010): 67–83,165–77.

[21] For an abstract of Tabata's paper, see Tomoji Tabata, "Using Random Forests to Identify Dickensian Style," abstract in "Language Individuation: A Symposium in Honour of John Burrows," University of Newcastle, New South Wales, Australia, 4 – 8 July 2011, online at http://www.newcastle.edu.au/Resources/Schools/Humanities%20and%20Social%20Science/Research/CLLC/2011%20conference/combined%20abstracts%20v3.pdf, pp. 3–4 (accessed 23 July 2012).

[22] Vickers, "Authorship Studies," 119.

[23] Patrick Juola, John Sofko, and Patrick Brennan, "A Prototype for Authorship Attribution Studies," *Literary and Linguistic Computing* 21 (2006): 169–78, esp. 174.

[24] Vickers, "Authorship Studies," 119.

[25] Vickers, "Authorship Studies," 119.

[26] Vickers, "Authorship Studies," 119.

Kinney frame their conclusions with a caution that Vickers often derides. No doubt he will claim that his matched *N*-grams are instances of the identical. But as he becomes more experienced, he will find that such moments of identity need not betoken the larger identity of shared authorship. While high levels of resemblance often reflect that larger identity, any authorial inferences formed about them can never amount to certainty. This last proposition reaches far beyond the gap between resemblance and identity that Vickers unwisely calls "the fundamental methodological weakness of computational stylistics."[27] The need to observe resemblances and differences, form inferences about identity, and act upon them—always without certainty—pervades our lives.

In the area of attribution, the matter of resemblance repays further thought. Without presuming to cover all possible cases, one may distinguish four relevant kinds of resemblance between texts. In attribution studies, the first two are usually of most interest. These two kinds present themselves in anonymous, pseudonymous, or disputed versions. Here, we must distinguish between single-author and collaborative work. The third kind is the imitation, whether parody, pastiche, or attempted forgery. The fourth is plagiarism. The first three kinds all offer broad, pervasive resemblances of style between the target text and the work, respectively, of its actual author, its actual authors, and its supposed author. The collaboration should show signs of two or more authors, each writing in his or her usual manner. The imitation is usually easy to identify because the imitator observes *and over-uses* some characteristic features of the target style while failing to register other such features. Plagiarism, the fourth kind, differs from the other three in that the culprit imports phrases, sentences, passages of the victim's work and either reproduces it verbatim or tries to disguise it. The plagiarism can therefore be expected to contain intermittent outbreaks of identity or near-identity to the work(s) from which it was pilfered. The evidence it affords differs, accordingly, from the broader but pervasive sort of resemblance arising in the other kinds of textual relationship that I have mentioned. Especially when the texts are tested in segments, as in Craig and Kinney, both PCA and Zeta are well able to distinguish among the four kinds of resemblance. To manage the like, Vickers would need to find a way of showing whether his chosen *N*-grams were huddled or dispersed. His current program, designed to detect evidence of plagiarism, may need to be adjusted.

While PCA is a long-established procedure, it was first used in literary studies, so I believe, when Nicholas McLaren of the Cambridge University Computing Laboratory recommended it to me in the early 1980s. He had already assured himself, by having me make trials with the *chi*-squared test, that

[27] Vickers, "Authorship Studies," 124.

my word counts offered results worth pursuing. Some years afterward, on the advice of the late Christopher Wallace of Monash University, we changed our way of using PCA and also took up Student's *t*-test to select whatever members of the word list were the most potent differentiae in a given case. These changes increased our usual levels of accuracy from less than 80% to more than 90%.[28]

With no direct experience of PCA, Vickers misses two crucial points. First, he rebukes Craig and Kinney for showing results in scatter plots representing only the first two PCs, thereby excluding all the valuable information lodged in the lesser components. (This is data reduction, as discussed above.) Vickers does not know what thirty years have taught us—that, in this sort of work as in many others, the PCs form a sharply declining hierarchy: each of them, that is, supplies less information than its predecessor. The third component can sometimes resolve a residual anomaly,[29] but I cannot recall a case where going any further meant faring any better. The governing principle, in plain terms, is concomitant variation. Specimens array themselves according to the most consistent affinities and disaffinities among the word variables. As each such array is garnered by PCA, it is set aside and the residual information is allowed to form a new array. Our experience has shown that, if differences of authorship, genre, and era are all in play, authorship usually prevails.[30] The most common exceptions arise from sharp differences in era or genre. Even these can often be resolved by the use of Student's *t*-test to select the most potent marker words. The heart of the matter is not the proportion of information gathered in the main arrays but the extent to which those arrays yield intelligible and verifiable patterns.

Vickers's second misconception of PCA is injudicious. He allows himself to be persuaded by a solitary correspondent that, after thirty years of work with PCA in literary studies, no user or critic has hitherto observed the simple fact (if fact it were) that the scores for the words highest of all in frequency engulf all others and rule the whole game. But fact it is not. As Craig plainly indicates,[31]

[28] A fully annotated sequence of the MINITAB commands for Wallace's application of PCA is set out in J. F. Burrows, "Not Unless You Ask Nicely: The Interpretative Nexus between Information and Analysis," *Literary and Linguistic Computing* 7 (1992): 91–109, esp. 103–4. In more recent versions of MINITAB, some modifications are required, but the underlying procedure is the same. PCA can also be run in other statistical packages.

[29] For a recent example, see Peter Anstey and John Burrows, "John Locke, Thomas Sydenham, and the Authorship of Two Medical Essays," *Electronic British Library Journal* (2009), article 3, http://www.bl.uk/eblj/2009articles/article3.html (accessed 30 November 2011).

[30] For a direct assessment, see Hugh Craig, "Is the Author Really Dead: An Empirical Study of Authorship in English Renaissance Drama," *Empirical Studies in the Arts* 18 (2000): 119–34.

[31] Craig and Kinney, "Glossary," s.v. "Principal Component Analysis (PCA)," in *Shakespeare, Computers*, 225.

PCs are not drawn directly from the table of frequency profiles but from a matrix of correlation coefficients derived from it. This, perhaps, is where Maciej Eder (Vickers's advisor on this point) went astray, with the consequences he describes.[32] Or perhaps he used covariance, not correlation. In any case, Vickers seizes it with enthusiasm.

PCA can fail when it is overloaded with data for too many different kinds of specimen texts. The Delta procedure is not affected in this way because each specimen in turn is compared with an unchanging base.[33] In the current version of Microsoft Excel, a single worksheet has the capacity to test over 3,000 separate specimens, authorial or otherwise, on a word list of over a million different word forms. The effective use of Delta is limited only by relevance and common sense.

Vickers, for once, allows himself to offer a half-concession: "It is true that, under certain favorable conditions, computing word frequencies can distinguish between two authors. But this is nothing new."[34] He then repeats the story of Mosteller and Wallace's groundbreaking work on the Federalist Papers as if to suggest that nothing has happened since. But, even in that very area, he is unaware of my use of cluster analysis to test the likely authorship of two large, pseudonymous sets of anti-Federalist Papers against the writings of many other pamphleteers.[35] As long ago as 1992, indeed, I used PCA to test a question of authorship in eighteenth-century fiction, comparing Tobias Smollett to fourteen of his contemporaries. The outcome supports the consensus among modern students of Smollett.[36] There has, no doubt, been less work on multicandidate

[32] Vickers, "Authorship Studies," 118n37.

[33] The Delta test is described in articles cited by Vickers; for fuller descriptions, see (all by John Burrows) "Questions of Authorship: Attribution and Beyond," *Computers and the Humanities* 37 (2003): 5–32; "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship," *Literary and Linguistic Computing* 17 (2002): 267–87; and "All the Way Through: Testing for Authorship in Different Frequency Strata," *Literary and Linguistic Computing* 22 (2007): 27–47. For assessments, see David L. Hoover, "Testing Burrows's Delta" and "Delta Prime?" in *Literary and Linguistic Computing* 19 (2004): 453–75, 477–95; Shlomo Argamon, "Interpreting Burrows's Delta: Geometric and Probabilistic Foundations," *Literary and Linguistic Computing* 23 (2008): 131–47; Peter W. H. Smith and W. Aldridge, "Improving Authorship Attribution: Optimizing Burrows' Delta Method," *Journal of Quantitative Linguistics18* (2011): 63–88; and Jan Rybicki and Maciej Eder, "Deeper Delta across Genres and Languages: Do We Really Need the Most Frequent Words?" in *Literary and Linguistic Computing* 26 (2011): 315–21.

[34] Vickers, "Authorship Studies," 116.

[35] John Burrows, "The Authorship of Two Sets of Anti-Federalist Papers: A Computational Approach," in *The Anti-Federalist Writings of the Melancton Smith Circle*, ed. Michael P. Zuckert and Derek A. Webb (Indianapolis, IN: Liberty Fund, 2009), 397–419.

[36] John F. Burrows, "Computers and the Study of Literature," in *Computers and Written Texts*, ed. Christopher Butler (Oxford: Blackwell, 1992), 169–204, esp. 175–80.

problems, where Delta is especially serviceable, than on direct contests between two candidates. But successful work with many candidates is nothing new.

With Delta, in particular, Vickers's dealings have been singularly ill starred. In his present review essay, he enters through a disused side door: "Although the Craig-Kinney collection does not disclose the fact, Burrows's methods have received serious criticisms and suggested reforms from other attributionists."[37] The criticisms offered in the three articles he cites all bear on Delta. A fourth, entitled "Deeper Delta," has since appeared.[38] Since Craig and Kinney make no use at all of Delta and never even mention it, they conceal nothing and have nothing whatever to disclose. As for the "serious criticisms" themselves, I believe that my own opinion will be widely shared by other attributionists: that procedures like Delta are always open to modification and refinement and that anyone who takes part in their development should be well pleased to see them optimized and kept in active use, not least by those who modify them.[39] Titles like "Testing Delta," "Optimizing Delta," and "Deeper Delta" are easy to live with. Shlomo Argamon's "Interpreting Burrows's Delta" offers a worthwhile critique; in a forthcoming review where he sees the field as "a mess," with "little to no general theory or deep understandings of the strengths and weaknesses of different methods," he acknowledges that "many significant accomplishments have been achieved" and regards machine-learning tools such as Delta as one of the "major approaches."[40]

Vickers's misadventures with Delta go back a little further. Shortly before his death in August 2007, Harold Love received top billing in the *Times Literary Supplement* (*TLS*) for a review of Vickers's book on the putative authorship of "A Lover's Complaint." Like all his work, Love's review is searching but genial. He accepts Vickers's case against Shakespeare's authorship but is not fully persuaded by the case for Davies of Hereford. He points out that, unless it is rigorously managed, the evidence of similarities in vocabulary is open to doubt. This leads him to recommend the use of "advanced kinds of statistical testing," including

[37] Vickers, "Authorship Studies," 115n27.

[38] Rybicki and Eder, "Deeper Delta."

[39] As mentioned earlier, the critics of Delta cited by Vickers are Hoover, Argamon, and Smith and Aldridge. In his paper at the Newcastle symposium on 23 July 2011, Hoover brought the matter up to date and, in a happy stroke, brought *N*-grams into play; see his abstract "Delta, Zeta, and Iota: An Ngrammatical Investigation," online at http://www.newcastle.edu.au/Resources/Schools/Humanities%20and%20Social%20Science/Research/CLLC/2011%20conference/combined%20abstracts%20v3.pdf, page 5 (accessed 30 November 2011).

[40] Shlomo Argamon, review of Kim Luyckx, *Scalability Issues in Authorship Attribution* (2010), *Literary and Linguistic Computing* 27 (2012): 95–97, esp. 95. For advance access, see http://llc.oxfordjournals.org/ content/early/2011/12/09/llc.fqr048.full (accessed 5 February 2012).

"John Burrows's Delta algorithm." But, in his view, Vickers's "brief comments" suggest that "he lacks real understanding of the field's present range and variety." He adds that "while Vickers is generous with numbers he is not their master."[41]

As Love expected, his review evoked a furious reply, which appeared only a week later and gave him keen amusement in a time of need. In his lengthy response, Vickers argued, "I am perfectly familiar with modern stylometry, but the Delta algorithm uses small vocabulary samples, such as the fifty most frequently occurring words."[42] But, as David Hoover shows in the very article where Vickers locates "serious criticisms," Delta gains in accuracy as the word list is extended from my initial 150 words to 600 or 700 words and more. As for my own contribution, Vickers seems to be convinced that I am still using very short word lists, as I did in my book on Jane Austen in the mid-1980s.[43] Nothing in his letter to the *TLS* or in his present review essay supports his claim that he is "perfectly familiar with modern stylometry."

So much for Vickers's attack on our methods. But, whatever testing methods are employed and whatever the data chosen, the question of sample size must be treated in a manner fit for the occasion. Until optimal levels are attained, larger samples of text usually yield more accurate results in tests of every kind: the statistical "law of large numbers" is their guarantee. Especially in poetry, however, many texts whose authorship is doubtful or unknown are very brief. With dramatic texts, especially when collaborative authorship is at issue, short scenes pose a similar difficulty. One answer is to treat texts below a given length as untestable. A second is to insert the brief disputed text into large, well-authenticated sets of work by each candidate for its authorship. The behavior of the set by the true author will be less affected than those by others. Fragile though it may be, the frequency profile of the disputed text will sit more easily among its brethren than in a gathering of strangers. A third answer, to be used later in the present essay, serves well with putative collaborations. When a disputed passage is left in its natural place within a series of "rolling segments," any stylistic fluctuations become visible. The best answer of all would lie in better tests than those we have. An improved capacity to test shorter samples at high levels of accuracy would be of more immediate value than further small improvements in treating longer texts.

To say such things, however, is to reject yet another of Vickers's attacks. Knowing that Craig and Kinney often break their texts into successive segments of 2,000 words each, he turns to the published abstract of a recent conference

[41] Harold Love, "Hallow the Shallow," *Times Literary Supplement*, 6 July 2007, 3–4, esp. 3.

[42] Brian Vickers, "'A Lover's Complaint,'" *Times Literary Supplement*, 13 July 2007, 6.

[43] J. F. Burrows, *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method* (Oxford: Clarendon Press, 1987).

paper by Maciej Eder.[44] From a long series of experiments, Eder concludes, among other things, that the results for samples of fewer than 3,000 words of English novels were "simply disastrous."[45] Instead of plunging forward, Vickers might have paused to look about him, not least in the very book he is reviewing. Many hundreds of the 2,000-word segments in Craig and Kinney's training sets, including those we have discussed, yield levels of accuracy well above 95%. Furthermore, their distinction between "Shakespeare" and "non-Shakespeare" is actually a distinction between almost 300 Shakespeare segments and more than 1,000 independent segments drawn from over 100 plays by some forty dramatists who were active between 1576 and 1642.[46] And, turning from the book he is reviewing to an article he cites,[47] Vickers might have noticed that, in a test of 200 Restoration poems, my level of accuracy for the longer texts is gratifying. For twenty poems of over 2,000 words, the true author ranked first out of twenty-five poets in nineteen cases and second in the twentieth. For twenty poems of 1,501 to 2,000 words in length, the true author ranked first in seventeen and between first and fifth in the other three. Nor does Vickers pause to observe that, albeit with samples larger than those we use, Eder achieves similar levels of accuracy. Jane Austen's estimable Miss Bates remarks, "What is before me, I see."[48] Vickers shares her gift when what is before him seems to disfavor his current opponent. His incapacity for seeing anything favorable is sublime.

Eder examined several corpora in several languages. The corpus requiring the largest sample size to meet his required level of accuracy was that of English novels, the one that Vickers singles out. But, chiefly by virtue of the marked difference between narrative and dialogic substyles, most novels are more mixed in style than almost any other texts. (The force of such differences and the need to take account of them can be seen in one of my earliest articles in this field.)[49] Since Eder's samples comprise random proportions of narrative and dialogue, any underlying stylistic habits would require larger than usual samples to make themselves evident. Eder recognizes this, but Vickers apparently does not. In their work on English dramatic texts, Craig and Kinney, not unnaturally, say nothing at all about the special sampling problems of prose fiction.

[44] Maciej Eder, "Does Size Matter? Authorship Attribution, Small Samples, Big Problem," online at http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/pdf/ab-744.pdf (accessed 30 November 2011).

[45] Eder, "Does Size Matter?" at page 2; see also Vickers, "Authorship Studies," 119.

[46] The plays are listed in Craig and Kinney's Appendix A, pp. 212–20.

[47] John Burrows, "Delta."

[48] Jane Austen, *Emma*, ed. R. W. Chapman, 3rd ed. (London: Oxford UP, 1933), 176.

[49] J. F. Burrows, "Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style," *Literary and Linguistic Computing* 2 (1987): 61–70.

III. Questions of Theory and the Choice of Data

It is now time to address topics where, even though Vickers seems not to think so, our work has more in common with his own. He claims, first, that computational stylistics lacks a theoretical foundation.[50] But anyone who uses stylistic evidence to approach questions of attribution is working within the ambit of Saussure's celebrated distinction between *langue* and *parole*, the language regarded as a system and the selection that any of its users makes. As students of literature rather than linguistics and as empirical problem-solvers rather than theoreticians, we seldom mention such generalities; but they are there and they can be seen as formative. On the question of whether a special theory of language need actually govern empirical inquiry, Harold Love stands with Vickers and is opposed by Willard McCarty.[51] In any event, the idea of distinguishable *paroles* is founded upon a broader belief in human individuality as expressed in many facets of our behavior.

Throughout his essay, Vickers sustains an analogy between language in operation and a woven fabric, and condemns us for being blind to what it implies. But consider the following passage. It is part of the preamble in my contribution to a volume from which Vickers quotes. (After digesting it, he might usefully read further and see how it is enacted.) Here is the passage:

> The real value of studying the common words rests on the fact that they constitute the underlying fabric of a text, a barely visible web that gives shape to whatever is being said. Despite a brave attempt made long ago (Bratley and Ross, 1981), we have yet to find satisfactory ways of tracing the interconnections of all the different threads, as each of the common words occurs and recurs in the full sequence of a text. We are able, nevertheless, to show that texts where certain of these threads are unusually prominent differ greatly from texts where other threads are given more use. An appropriate analogy, perhaps, is with the contrast between handwoven rugs where the russet tones predominate and those where they give way to the greens and blues. The principal point of interest is neither a single stitch, a single thread, nor even a single color but the overall effect. Such effects are best seen, moreover, when different pieces are put side by side. That, at all events, is the case I shall present.[52]

If that is really our belief, Vickers might respond, why do we destroy the fabric by picking out some of its components? The idea that classifying and counting

[50] Vickers, "Authorship Studies," 114–17.

[51] Harold Love, *Attributing Authorship: An Introduction* (Cambridge: Cambridge UP, 2002), 132–62; and Willard McCarty, *Humanities Computing* (Basingstoke, UK: Palgrave Macmillan, 2005), 139–55.

[52] John Burrows, "Textual Analysis," in *A Companion to Digital Humanities*, ed. Susan Schreibman, Ray Siemens, and John Unsworth (Oxford: Blackwell, 2004), 323–47, esp. 323–24.

are destructive is a primitive fallacy. We pick out words for precisely the same reason that he picks out *N*-grams. It serves the purposes of analysis.

Throughout its long and reputable history in many fields of inquiry, analysis has usually depended upon comparison, whether between specimen and specimen or between each specimen and an acceptable base. In some cases, the object is to identify features *peculiar to a given class* of specimens (such as texts by Shakespeare), in others to identify features that *strongly preponderate or are markedly deficient in that class*. The latter approach often uses statistical methods in order to distinguish significant differences from chance effects. The former, usually arithmophobic, approach appears transparent but is perfectly open to abuse. Neither approach is of any merit unless it rests upon equitable comparisons.

The simple elements of written language, in increasing order of complexity, are alphabetic characters, syllables, words, phrases, clauses, sentences, and paragraphs. A case can also be made for including punctuation marks, numerals, and word spaces. From another perspective, the chosen constituents (although not quite simple elements) of a text might include rhymes, alliterations, tropes, Latinate word forms, or many other sorts of identifiable feature. In dealing with questions of attribution, I believe, the object is clear: we seek to identify a body of idiosyncratic phenomena, whatever their nature. Where then is such analysis best directed?[53]

Vickers is right to claim that *N*-grams retain something of the forward movement of a text but not to pretend that single words are devoid of that capacity. Every single conjunction, relative pronoun, and preposition bears implications

[53] Among the simplest elements of language, words have received most attention from attributionists in recent years. But in an article published in 1996, Richard Forsyth and David Holmes showed that when sequential strings of alphabetic letters were counted as entities, the results were more accurate than those then being yielded by the very frequent words. These strings, in which interword spaces were left in place, also gave better results than digrams, digraphs, or single letters. Over their range of tests, not all authorial, STRINGS yielded almost 80% accuracy, LETTERS only 69%, with the other three bunched below 75%. But, as the authors point out, their STRINGS, unlike their other data, had to be chosen in advance as those that best distinguished between their training sets. My colleagues and I were much impressed by this piece of work and expected to hear more of it. We did not turn that way ourselves because, by that time, we were achieving about 90% accuracy with WORDS by using Student's *t*-test to select the best "markers." In the event, this use of STRINGS did not go far beyond that article. Extraneous obstacles have stood in Forsyth's way and Holmes has reverted to WORDS in his more recent work. In their most recent article, Holmes and Daniel Crofts use PCA (supported here by cluster analysis and Delta) and refer to it as "the first port-of-call for attributional problems" (186); see David I. Holmes and Daniel W. Crofts, "*The Diary of a Public Man*: A Case Study in Traditional and Non-Traditional Authorship Attribution," *Literary and Linguistic Computing* 25 (2010): 179–97.

about its antecedents and its more or less immediate successors. The articles and the personal pronouns are more local, but not lacking, in such effects. A noun requires a verb or a preposition. Nouns and verbs tend to gather adjectives and adverbs. By setting aside (without ever forgetting) the sequentiality of words in context, we shift away from that familiar perspective and treat words as members of their respective grammatical classes. Comparing frequencies enables us to assess the weight attaching to the various classes in texts of different kinds or different authorship.

Vickers so relishes Juola's neat jibe at our "bag of words"[54] that he keeps coming back to it. But are we, like Autolycus, mere snappers up of unconsidered trifles? The illustrations to be offered later in this essay show that what comes out of the bag is not without intelligible shape. As for what goes in, a table of standardized frequency profiles for a suitable range of texts has more in common with the contents of a well-ordered filing cabinet than with those of your average handbag. I call to mind a prescient passage from *Macbeth*, as true of words as of either dogs or men, and, renouncing any claim to the title conferred on me by Vickers, I hereby declare William Shakespeare the great progenitor, the onlie begetter, of computational stylistics:

> the valued file
> Distinguishes the swift, the slow, the subtle,
> The House-keeper, the Hunter, every one
> According to the gift which bounteous Nature
> Hath in him clos'd.
> (*Macbeth*, 3.1.94–98)

I originally turned to computers and statistics because I had noticed, as a reader, that Jane Austen's characters differ markedly from each other in their recourse to many rather commonplace words. As I delved among increasingly frequent words without any apparent lessening of these differences, it became clear that card-indexing was inadequate for marshaling such observations. Computer assistance revealed, much to my surprise, that the differences still prevailed even among the most frequent words of all, words that have been described as verbal sludge. Taken all together, even a small number of them embrace a large proportion of any ordinary English text. They are, of course, almost all definable as function words rather than lexical words. Their function is highlighted when each of them is counted with its own kind. Along with others not much less frequent and in a manner scarcely perceptible even to a sensitive reader, the function words provide the main strands of the whole fabric. In the "slot-and-filler"

[54] Patrick Juola, "Authorship Attribution," *Foundations and Trends in Information Retrieval* 1 (2008): 233–334, esp. 253.

model proposed by scholars of whom Vickers happens to approve,[55] function words form the slots and lexical words the fillers.

My original principle of selection, from which we have rarely diverged, was that, in any of my inquiries, the texts themselves should supply the word list according to their own descending order of frequency. That nullifies the risk of choosing just such data as will favor a desired outcome. Three later modifications of the principle should be noted. First, as has been said, we now (when it is appropriate) use our training texts to identify marker words. Second, while still allowing their training sets to determine the selection, Craig and Kinney decided to form two such lists consisting, respectively, of function words and lexical words. Their object was to obtain independent data for corroborative testing. Their success, *pace* Vickers, is demonstrated by their results. The third and largest change lies in the gradual extension of the word lists. When I began to enlarge my original small set of words, the Minitab statistical package[56] had a ceiling of ninety-nine variables for the correlation procedure used in PCA. My publications in the late 1980s and 1990s observe that constraint. Given the use of marker words, PCA operates effectively within that limit. But the development of Delta changed the game. Its much simpler arithmetical basis made it possible to use Microsoft Excel instead of MINITAB and to achieve a vast increase in capacity. My first extension to 250 words gave good results but David Hoover was soon to show an ever increasing accuracy when the 600, 700, or 800 most frequent words were brought into account. Given texts of sufficient length, Hoover now finds that a list of 2,000 words works well. In their recent article, Rybicki and Eder go even further.[57] Where PCA is at its best with a comparatively small number of strong variables, Delta has a most voracious appetite. This makes it ever more possible to benefit from the multiplicity of weak determinants favored in some branches of scientific inquiry.

For my purposes, however, Delta is better when it is a little underfed. Many texts of doubtful authorship are too short to approach with long word lists. Operating as we are at high levels of accuracy, we cannot benefit as much from marginal improvements as from corroborative testing. And the power of tests like Zeta and Iota is reduced when Delta is allowed to invade their frequency strata: when that is allowed, the effect is either to weaken them by absorbing some of "their" words or to compromise their independence by using some of the same words in different tests.

[55] Vickers, "Authorship Studies," 135.
[56] Minitab Statistical Software, http://www.minitab.com/.
[57] Hoover, "Testing Burrows's Delta"; and Rybicki and Eder, "Deeper Delta."

To round off this brief account of our choice of data, I raise two further matters. If Vickers were indeed as familiar with modern stylometry as he claims or had given due attention to the work he cites, he should know that some of the methods now in use begin by embracing *all* the words of the texts under examination and then sift through them in much the same fashion as he should sift his *N*-grams.[58] The underlying logic of the Iota test is just as his should be. From two comparable bodies of text for training sets, it identifies all those words that occur in each one but not in the other. These two long lists are then matched against the target text. The postulate is that an author who uses a particular word even once is rather more likely to return to it than an author who has not used it is likely to take it up. The fact that this test yields extremely high levels of accuracy supports the idea that, whether we know it or not, we have our preferences in fillers, as well as slots. They, too, make up part of our personal *paroles*.

I take up the second of these matters at the suggestion of an anonymous reviewer of this essay who suggested that I should comment on the admission of subject-specific lexical words among the more author-oriented function words, especially in longer word lists. He offered the ships and whales of *Moby Dick* as an example. Let me first respond by acknowledging the difficulty the reviewer identifies. Such words would not help to identify "Bartleby, the Scrivener" as Melville's; the problem is not confined to lists based on single texts. In the work of all the Cambridge Platonists of the late seventeenth century, "God" and "Christ" are located around the middle of the hundred most frequent words of all. In the verse of their contemporaries, the obscene wits of Rochester's circle, such words are displaced by quite another set of monosyllables. Yet even here, persistent authorial preferences can be observed as when some of them prefer the archaic verb *swyve* to its more familiar synonym. And, even in the more temperate linguistic climate of Renaissance drama, Craig and Kinney find that *gentle* occurs more than twice as consistently, *brave* only half as consistently, across the range of Shakespeare's work as across that of his contemporaries.[59]

When any word list is simply extended down the frequency hierarchy, the proportion of lexical words increases rapidly. In the list used in part IV of this essay, based on Craig's set of 202 plays, the top 100 words of the list include only 9 lexical words. In the stratum between the 401st and the 500th, there are 90 such words. Those of the top stratum are commonplace—*good* and *well*, *see* and *say*, *make* and *take*, and so forth. In the words at the 490th stratum, we find such more volatile words as *soft*, *bloody*, *dream*, *fools*, *health*, and *laugh*.

[58] See, for example, Cyril Labbé and Dominique Labbé, "A Tool for Literary Studies: Intertextual Distance and Tree Classification," *Literary and Linguistic Computing* 21 (2006): 311–26; and John Burrows, "All the Way Through."

[59] Craig and Kinney, "Methods," 16–18.

There is no doubt that broadly based word lists and truncated word lists alleviate the difficulties that may flow from the inclusion of too many subject-specific words in any test of authorship. But, as we have just seen, a broadly based list soon gathers in words that may well be subject specific. With truncation, a sort of rigor is obtained by the sacrifice of potentially valuable evidence of authorship: in the aforementioned list, *brave* and *gentle* rank 298th and 325th, respectively.

How, then, should we manage a mixed flock of authorial sheep and subject-specific goats? Not, I would insist, by picking and choosing and thus allowing our prejudices to run wild. If there is to be any culling of the word list, it needs to be according to predetermined rules. For his very long word lists, David Hoover culls such words as occur disproportionately often in any one of his specimen texts. I have recently been experimenting successfully with the exclusion of all words that do not occur in every training text. Some true authorial idiosyncrasies are lost but resemblances or differences of subject matter no longer affect the issue. The corroborative use of independent tests is, as always, of great value. But the best line of approach to these difficulties is to form appropriate training sets and conduct suitable advance trials before addressing the problem text. In part IV of the present essay, for example, the results for *3 Henry VI* are validated in this way. Without presuming to encompass Shakespeare's style, one may identify idiosyncratic features of his *parole* and make good use of them.

Turning finally, from our use of word frequencies to the use of *N*-grams, I observe that, when it comes to his own work, Vickers is at ease with a loose-fitting definition. He speaks of "phrasal repetends, collocations, *N*-grams, call them what you will."[60] This enables him to dress many of them in borrowed robes. Let us try to be more exact and call them what they are. Collocations and *N*-grams are general terms for the phenomena in question. The former emphasizes the proximity, within a specified distance, of the words to be regarded as an entity; the latter generalizes the number of strictly consecutive words to be embraced. Phrasal repetends, however, are a special subgroup.[61] They are phrases of a kind whose behavior led Ian Lancashire toward the field of cognitive linguistics where he has contributed so much. By lumping them all together, Vickers gives a free ride to collocations of less force. The fact that most collocations are not repetends is only to be expected. But he attaches a cognitive weight to any collocation that can be seen as peculiar to his favored authorial candidate. The difficulty is that, repeated or not, peculiar or not, many collocations are only

[60] Vickers, "Authorship Studies," 138.

[61] See Ian Lancashire, "Phrasal Repetends in Literary Stylistics," in *Research in Humanities Computing 4* (see n. 9 above), 34–68.

quasi-phrases and many others are gibberish. Can such phenomena seriously be said to have been *formed* or to make part of an author's "phraseognomy?" A memorable passage from *Macbeth* offers sufficient illustration:

> Will all great *Neptunes* Ocean wash this blood
> Cleane from my Hand? no: this my Hand will rather
> The multitudinous Seas incarnadine,
> Making the Greene one, Red.
>
> (*Macbeth*, 2.2.60–63)

Masquerading for a happy moment as a reader-scholar, one might remark the astounding lexical choices we have here; the strong opening hyperbole; the surging onomatopoeia of the third line set against the monosyllabic thudding of the fourth; and, above all, the power of the passage as one dark step downward on the inexorable path from "A little Water cleares us of this deed" (2.2.67) to "Yet heere's a spot" (5.1.30). But our business is with attribution and with seeking collocations. Vickers now tells us that we must not hope for Arnoldian "'touchstones.'"[62] This is well advised. Any collocation embracing "incarnadine" is unmatched in Shakespeare and cannot serve as a repetend, phrasal or otherwise. There are several matches for "multitudinous," but Jonson and Dekker also figure there. "Neptune" is almost anybody's word. "This my hand" and "wash this blood" deserve investigation. After that, we are left, alas, with collocations less redolent of high tragedy than of golf, snooker, and shopping—making the green, making the . . . red, the green one, and so forth. Of these, moreover, only "making the green" has much claim as a cognitive entity. With or without the Folio's comma, discarded by modern editors to admit a stronger reading, it would be specious to suggest that Shakespeare wrote "making the . . . red" *as a phrase*. His business was with a double phrase whose latter part contains an implied preposition: either "making the green one, [into] red" or "making the green [into] one red." In either reading, that is what Lancashire would call a "chunk," the sort of entity the mind does form. As for gibberish, "hand no this" and "green one red" will do. Such collocations are easily matched: "I am holding one green, one red, and two purple." "In that hand?" "No, this." But I find no Shakespearean cognition in either of them and never a whiff of syntax.

Although his explanations and his examples are often at odds, it appears that Vickers is dealing principally in trigrams: "When two texts are 'read' in parallel, the program automatically identifies every instance where the same three consecutive words occur in both."[63] That is just one small step beyond the digrams

[62] Vickers, "Authorship Studies," 142.
[63] Vickers, "Authorship Studies," 139.

tested by Forsyth and Holmes in 1996. However, Vickers accepts expansion beyond the trigram when the texts allow it.

Another area of uncertainty surrounds the laudable proposition that "authorship attribution studies should always move from the known to the unknown."[64] Everything he says here about the Additions to *The Spanish Tragedy* and about his work on Thomas Kyd has him traveling in the opposite direction. He proceeds by identifying the successive trigrams in his *dubium* (the Additions, for example), selecting those shared with *a chosen author*, and then seeing if they can be matched elsewhere. The near-simultaneity of these steps does not rectify their sequence. Vickers lays weight on what he calls "negative checks,"[65] but what he describes is only a filter. This is not sufficient to his purpose, which actually obliges him to give each eligible candidate a turn.

In his conclusion, he makes much of the parallels that he and Warren Stevenson find between the Additions to *The Spanish Tragedy* and Shakespeare's plays. It is invidious for anyone with our current technical advantages to complain that earlier scholars like Stevenson did not undertake a sufficient range of comparisons. But Vickers has no excuse. After searching in a set of sixty-four plays dated before 1596, he announces that he has found many collocations from the Additions that are matched only in Shakespeare. No matter how numerous, such collocations carry little or no evidential weight until rational questions like "How many should be expected?" and "How many others are there in work by other true candidates?" are entertained. Of the dramatists active before 1596, only Shakespeare is regarded as a possible author of the Additions. Of plays dated before 1596, Harbage and Schoenbaum's *Annals* gives Shakespeare as sole author of nine. Since his plays tend to be longer than those of most other authors, they should contribute more than twice as many words to Vickers's database as the most prolific of the other early dramatists, except for John Lyly, who has eight plays. It follows that (Lyly apart) Shakespeare must provide Vickers with more than twice as many unique matches as any of those others before the numbers we are offered mean anything at all.

As things stand, over a quarter of the plays in Vickers's set of sixty-four are the work of dramatists who were dead long before the time of the Additions. When he decides to show how often the Additions contain unique matches for Dekker, Jonson, and Webster, he may begin to have a case. Even then, he will need to find a way of balancing the several corpora to allow for differences in size. It seems likely that Shakespeare will finally prevail. However, given the inconclusiveness of the external evidence, the only proper support for that opinion lies in Hugh Craig's scrupulous comparisons and his strenuous attempts

[64] Vickers, "Authorship Studies," 139.
[65] Vickers, "Authorship Studies," 141.

to challenge his own findings. Craig alone gives each of the known candidates a full hearing and tests for the possible role of other candidates.[66] Toward the end of his review essay, where he inaugurates a bright new era in authorial studies, Vickers seems to envisage himself as "the bird of lowdest lay, / On the sole *Arabian* tree" ("The Phoenix and the Turtle," ll. 1–2). Those of us who have read widely and attentively in the field of attribution studies know that the new era began some time ago, and that it owes much to many more scholars than the handful I have mentioned. We shall await his eventual contribution with interest. But only when he makes strict, appropriate, and equitable comparisons will he arrive at the beginning of the beginning.

IV. A Set of Results: By Their Fruits. . . .

In this last phase of my argument, I turn from discussion to demonstration. Attributional scholarship seeks to distinguish the work of one author from that of one or more others in order to identify the author of a disputed or an anonymous text. Vickers claims that the study of *N*-grams yields much better results than the study of word frequencies. I respond that this question must remain open until he puts his *N*-grams to proper use. It is incumbent on me, nevertheless, to show that word frequencies do offer serviceable evidence. In the following examples, my choice of data and procedures directly belies some of the misrepresentations I have described. As to the outcome, Shakespeare's work is clearly distinguished from that of his fellows. (In my Figure 4, for example, every member of each little army stands on the appropriate side of no-man's land while two plays thought to be of mixed authorship stand between them.) Shakespeare's work also stands apart from that of other dramatists when we analyze groups of plays (Figures 2 and 3) and when we break plays into segments (Figures 5 to 8). It should be observed that, while the evidence of this section rests entirely upon word frequencies, it does not directly testify to the value of Craig and Kinney's findings. But perhaps it will encourage the belief that they deserve a better hearing than Vickers gives them.

The test employed here is Delta, which allows comparisons among many specimens, authorial or otherwise, at one time; which admits the use of extensive word lists; which offers high levels of accuracy with samples of as few as 2,000 words; and which can be used to give a helpful impression of even shorter passages. All four of these statements belie Vickers's claims to the contrary.

The word list chosen for my first three graphs is a descending hierarchy of the 500 most frequent words in Craig's machine-readable collection of 202 plays

[66] Hugh Craig, "The 1602 Additions to *The Spanish Tragedy*," in *Shakespeare, Computers*, 162–80.

dated from 1576 to 1642. For the remaining graphs, which deal with shorter samples, the same list is truncated to its 300 most frequent words. In both cases, this list provides the set of word variables from which corresponding frequency profiles for the chosen specimens are compiled. Each profile is compared with the same base. Since each comparison is independent of the rest and since the frequencies for each specimen are standardized in proportion to its word length, Delta is more robust than many other tests and is not much affected when the specimens differ in length. The standardization prevents the words at the head of the list from dominating the outcome.

With each successive specimen, Delta calculates the absolute divergence of the score for each word variable from that of the target text. The average divergence for each specimen is its Delta score. The lowest Delta score among all the specimens represents the least overall divergence from the target, identifying that specimen as "least unlike" and ranking it first of them all. The others rank after it in ascending order of their Delta scores.

The present set of specimens comprises sixty-six well-authenticated single-author plays and two putative collaborations. Twenty-eight of the sixty-six are Shakespeare's, sixteen dated before 1600 and twelve from 1600 onward. Thirty-eight are by fourteen other dramatists, and are all dated before 1600. The other two plays, both from the early 1590s, are *Titus Andronicus* and *3 Henry VI*—one widely accepted as collaborative, the other thought to be so. The dating used is that of the 1964 edition of Harbage and Schoenbaum's *Annals of English Drama 975–1700*.[67]

For my first pair of graphs, which differ from each other by a significant change of base, I worked with a broad brush. The thirty-eight non-Shakespeare plays were gathered in their fourteen authorial sets. The two collaborative pieces were left to stand alone. But the two groups of Shakespeare plays were used in such a way as to let them change roles. The twelve later plays (Figure 2) provided the base against which all the rest are tested. But instead of taking the sixteen earlier Shakespeare plays as one authorial set, I separated them into four subsets, each containing four plays. The immediate effects were to produce four separate Delta scores for Shakespeare and to bring him into a more equitable comparison with playwrights who wrote so much less than he. (In Figure 3, the sixteen earlier plays provide the base and the twelve later plays form three subsets of four plays.) The first four subsets were composed by arranging the sixteen earlier plays in the alphabetical order of their titles. The first, fifth, ninth, and thirteenth were gathered as "Shakespeare A"; the second, sixth, tenth, and

[67] Alfred Harbage, *Annals of English Drama 975–1700*, 2nd ed., rev. S. Schoenbaum (Philadelphia: U of Pennsylvania P, 1964).

fourteenth as "Shakespeare B"; and so on. For Figure 3, the three subsets drawn from the twelve later plays were gathered in the same fashion. In terms of dates and genres, the process yielded a random mixture of Shakespeare's plays within two large chronological groups.

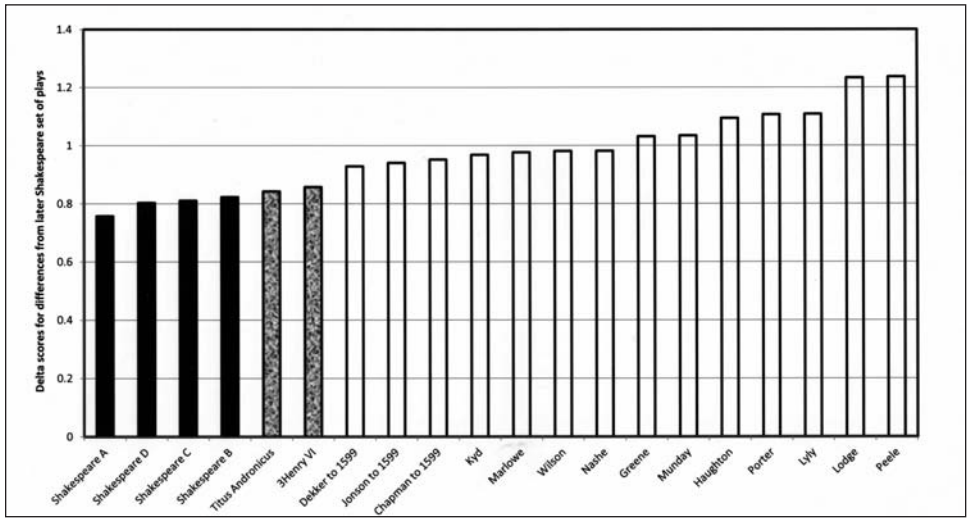

Figure 2: Fifteen Elizabethan dramatists and two putative collaborations, showing levels of difference from later Shakespeare. The Delta scores are based on the 500 most frequent words of 202 English plays, 1576–1642.

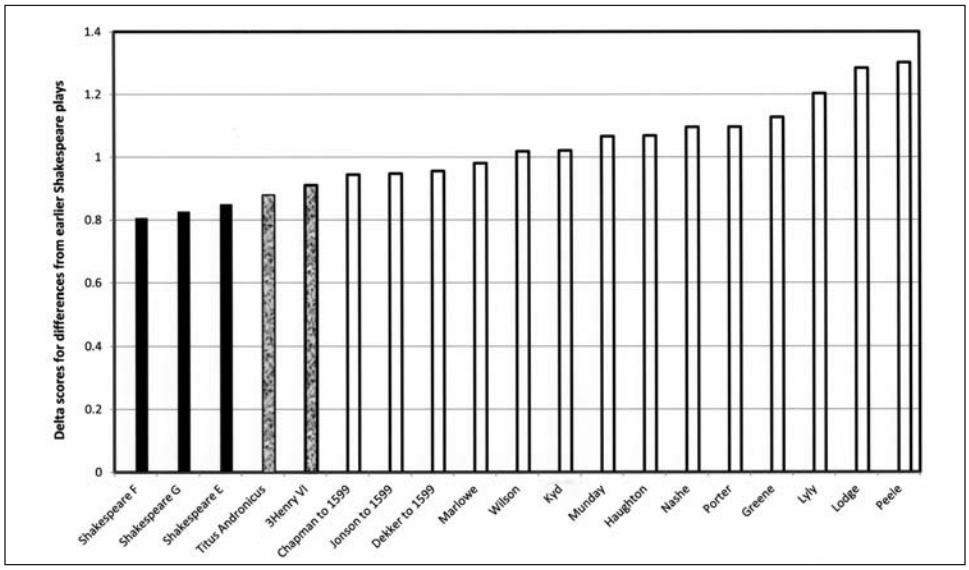

Figure 3: Fifteen Elizabethan and early Jacobean dramatists and two putative collaborations, showing levels of difference from earlier Shakespeare. Delta scores are based on the 500 most frequent words of 202 English plays, 1576–1642.

Figure 2, which is based on the 500 most frequent words in Craig's full set of plays, shows Delta scores in ascending order for eighteen authorial groups of plays and for the two collaborative texts. The four subsets comprising Shakespeare's earlier plays yield the lowest Delta scores; the two collaborative texts rank next above them. Since these six entries diverge least of all from the mean-scores for Shakespeare's later plays, the effect of authorship prevails.

Figure 3 shows an essentially similar outcome. The difference is that Shakespeare is now put at a chronological disadvantage: his late work and the work of his early peers are now assessed for their divergence from his early work. Evidence presented elsewhere by Hugh Craig shows overall changes in the language of drama in the period from 1580 to 1615 (and beyond).[68] This sits well with the belief of language historians that the period was marked by great changes in the language generally. But these changes do not match the force of authorship as a differentia. In each graph of the present pair, the Shakespeare entries stand apart from all the rest; the plays in which he is thought to have collaborated occupy an intermediate position, and the rest range out beyond them. The order of entries for the other fourteen dramatists differs a little between graph and graph. There is no clear difference between the single-play entries and those derived from larger corpora.

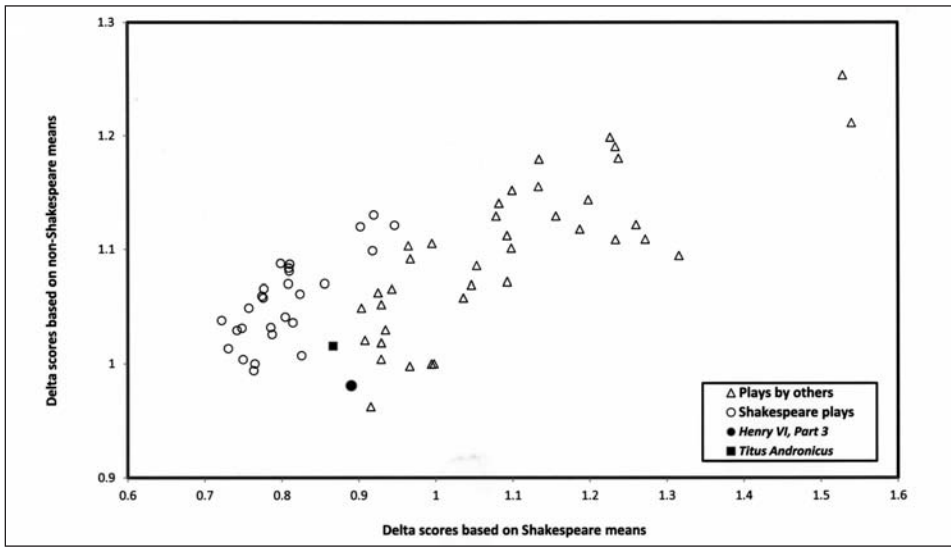

Figure 4: Sixty-eight plays, with Delta scores for each play as measured against mean scores of the two main groups of plays. Scores based on the 500 most frequent words of 202 English plays, 1576–1642.

[68] Hugh Craig, "Grammatical Modality in English plays from the 1580s to the 1640s," *English Literary Renaissance* 30 (2000): 32–54; and Hugh Craig, "*A* and *an* in English Plays, 1580–1639," *Texas Studies in Literature and Language* 53 (2011): 273–93.

Figure 4 takes the sixty-eight plays as separate specimens and arrays them in a scatter plot. On the horizontal axis, the Delta score for each specimen represents its overall divergence from the mean scores of the twenty-eight Shakespeare plays. The vertical axis shows divergences from the mean scores of the thirty-eight non-Shakespeare plays. Each play, accordingly, has a "Shakespeare" and a "non-Shakespeare" score. The two collaborations continue as independent entities.

The two main clusters stand apart, separated by an imaginary straight line slanting upward from left to right. The Shakespeare plays form much the more tightly knit group and Student's *t*-test assesses the probability that the two groups are members of the same population as just 1 in 10 million million—10 old-fashioned British billions. The diffuseness of the non-Shakespeare group is expected because it embraces plays by fourteen dramatists. The horizontal axis, accordingly, is chiefly responsible for separating group from group. On this measure, as on others, the Shakespeare plays that emerge as outliers are the subgroup of four located above and to the right of his main group—*Coriolanus*, *The Merry Wives of Windsor*, and the two earliest comedies. It should also be noted that these four entries lie near the middle of the vertical axis—atypical for Shakespeare but with no evident claim to be the work of other hands.

As might reasonably be expected, *Titus Andronicus* and *3 Henry VI* lie between the Shakespeare and non-Shakespeare sets. Their nearest neighbors to the left are *Romeo and Juliet* and *Henry V*. To the right are Peele's *Edward I*, Greene's *James IV*, and Marlowe's *Edward II*. Other early plays, mostly histories, predominate in the bottom left-hand corner of the graph.

Each of my remaining graphs focuses on a particular play, showing how its behavior varies as its dramatic action goes forward. In these graphs, the play being tested is broken into successive rolling segments of 2,000 words, each of which discards the first 200 words of its predecessor and advances by 200 words. The pattern is like that which is used in "rolling quarterly periods" in reports of many kinds: January to March, February to April, and so on. Because the segments are of no great length, the previous word list of 500 words is truncated to its top 300. Each segment in turn is treated as the target text and compared with each of the other sixty-seven plays we have been studying. The fluctuating scores across the range of segments show their varying affinity for any, many, or all of these other plays. While the broad picture matters most, there is a useful side effect in that the fluctuations give some insight into the behavior of phases, scenes, passages too brief to be tested as entities.

In Figure 5, *Romeo and Juliet* is tested in this fashion, singled out from Figure 4 for its proximity to the two collaborative plays. The successive segments range across the horizontal axis and their scores are represented on the vertical axis. To

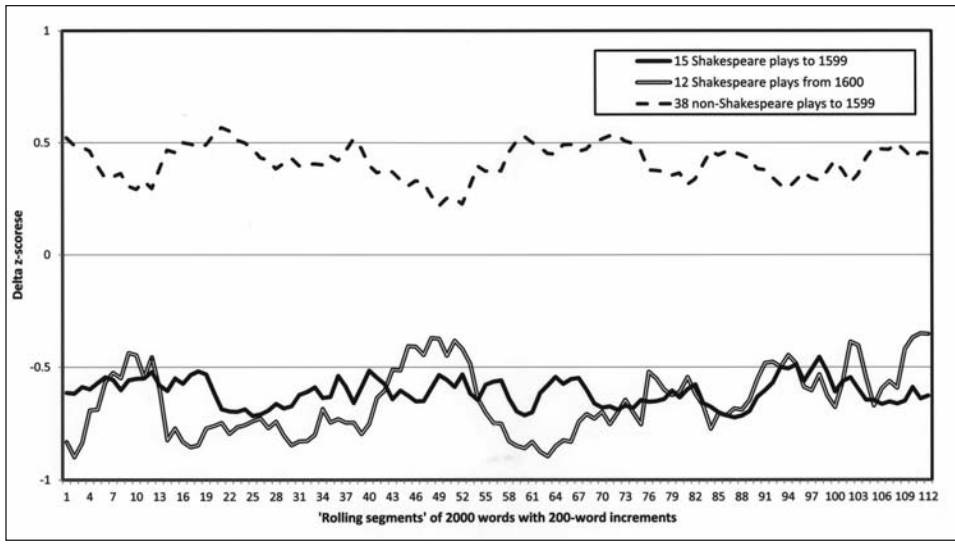

Figure 5: *Romeo and Juliet*. Delta *z*-scores for Shakespeare and fourteen early dramatists. Scores based on the 300 most frequent words of 202 English plays, 1576–1642.

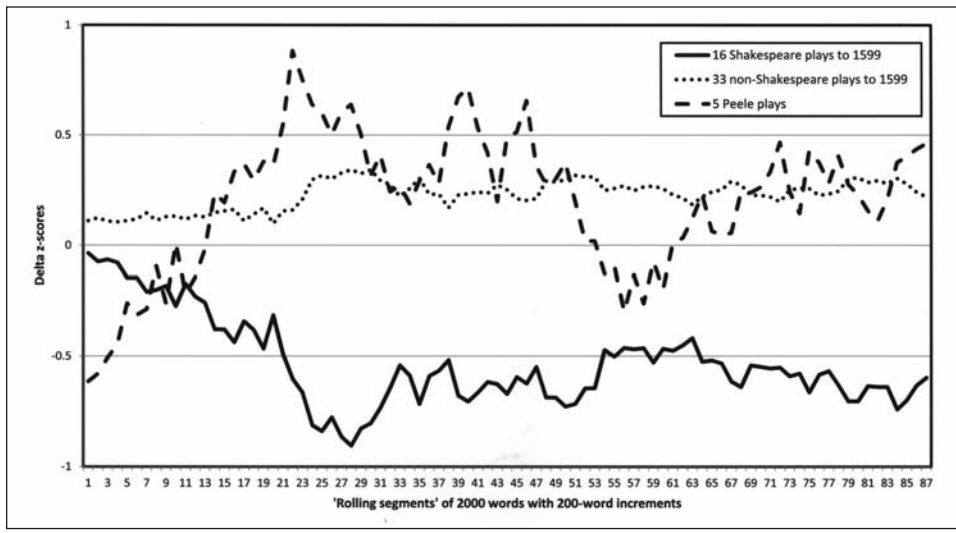

Figure 6: *Titus Andronicus*. Delta *z*-scores for Shakespeare, Peele, and thirteen other early dramatists. Scores are based on the 300 most frequent words of 202 English plays, 1576–1642.

allow proper comparisons between one graph and another in this group of four, the scores are given as Delta *z*-scores. These are based on means and standard deviations in the Delta scores for the sixty-five single-author plays that remain when the current target play is set aside for testing.

At no point does the entry line for the non-Shakespeare plays come within reach of either the earlier Shakespeare or the later. These two, meanwhile, never move more than half a standard deviation apart from each other. Most of Shakespeare's other plays yield equally clear-cut results when they are tested in this way. In some early plays, a few errors do occur. These, I believe, are the effect of his not yet having settled firmly into his style. (The last graph in this present series is a case in point.) If the authorship of *Romeo and Juliet* were called into question, Figure 5 offers 112 little affirmations of Shakespeare's claim. These would need to be compared, in the same fashion, with the claims of every other candidate.

Figure 6 shows how that can be done. *Titus Andronicus* offers an excellent test case. There is a consensus that the play is collaborative, that George Peele was Shakespeare's partner in the work, and that Peele contributed four scenes (1.1, 2.1, 2.2, and 4.1). While the consensus view is not a known truth, the evidence supporting it is strong and wide ranging.[69] To try the case again with a different test is not so much to offer corroboration as to assess the efficacy of the test itself.

Only the recognized claimants are supported by Figure 6. The entry line for the other early dramatists shows no affinity for *Titus Andronicus*. But Peele stakes a claim at the beginning and again around segments 51 to 60. Shakespeare predominates elsewhere. All that fits well with the consensus. The first three scenes, all attributed to Peele, run to just over 5,000 successive words, and his last, 4.1, to just over 1,000. Now our first segment embraces the first 2,000 words; by Segment 11, we have 4,000; and by segment 16, we have incorporated the remainder of Peele's first part. The weakening of Peele's affinity and the strengthening of Shakespeare's begins a little sooner than this would suggest; but it is not until after segment 26, where Peele's first part drops out completely, that the affinity for Shakespeare comes into full force. When Peele reappears, on cue at around segment 51, it is with less than his earlier force. That is because the 1,030 words of 4.1 can never yield much more than half a segment. The segments in which it participates range from nearly half Shakespeare to much more. It is in effects like this that rolling segments enable us to study the behavior of comparatively short scenes although not the very short ones.

[69] The case is set out fully in Brian Vickers, *Shakespeare Co-Author: A Historical Study of Five Collaborative Plays* (Oxford: Oxford UP, 2002), 148–243, and Table 3.11.

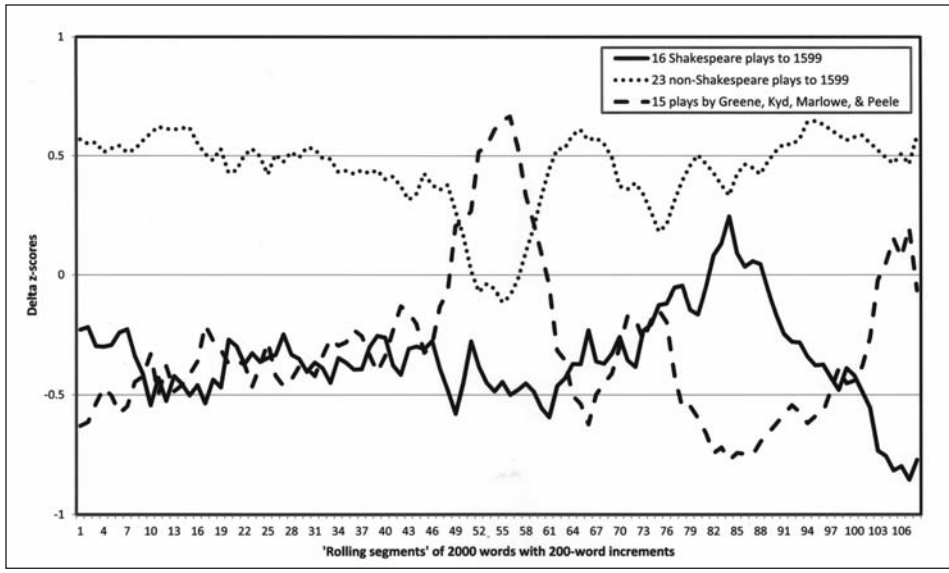

Figure 7: *3 Henry VI*. Delta *z*-scores for Shakespeare, four early collaborators, and ten other dramatists. Scores based on the 300 most frequent words of 202 English plays, 1576–1642.

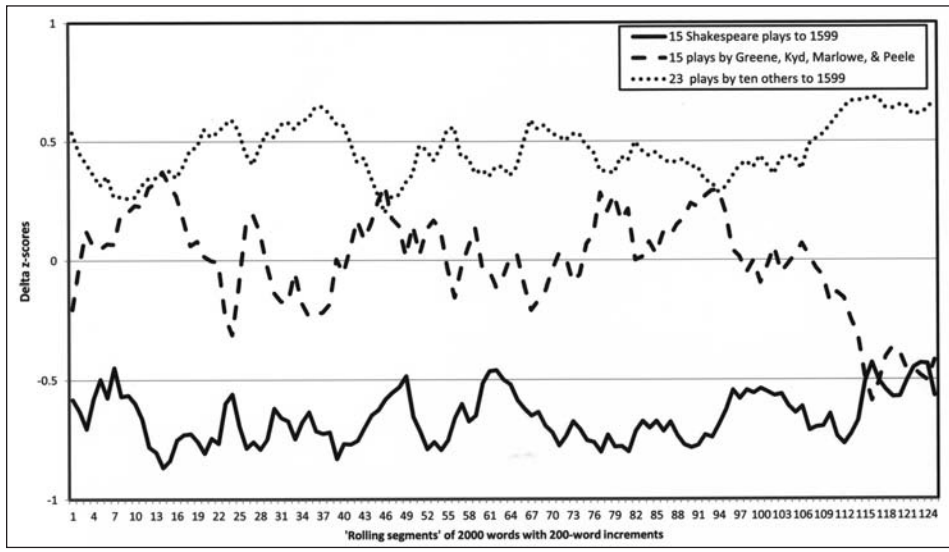

Figure 8: *Richard III*. Delta *z*-scores for Shakespeare, four early collaborators, and ten other dramatists. Scores based on the 300 most frequent words of 202 English plays, 1576–1642.

Figure 7 travels into less well-charted waters. Most modern Shakespeare scholars regard 3 *Henry VI* as a collaboration in which Shakespeare participated. Greene, Marlowe, and Peele have all been proposed as possible collaborators. At that early date, the possibility of Kyd's involvement should also be entertained. Although there is broad agreement so far, it does not extend to the question of who wrote which parts of the play. In a recent study,[70] Hugh Craig and I tried to make a little progress. Our evidence supports the view that the play is collaborative and that the only likely contributors are those named above. It favors Shakespeare's authorship of scenes from the middle and the end of the play but not of some others, especially in Acts 1 and 4. The outcome, in short, was much as it is represented in Figure 7, which approaches the matter in the same way but uses quite a different word list and lacks the corroborative support of the other tests used in our study of this play.

Figure 8 makes a natural complement to Figure 7. It approaches *Richard III* in exactly the same way. The ten "other dramatists" gain no support. Shakespeare stands unchallenged until near the end where the entry line for Greene, Kyd, Marlowe, and Peele suddenly turns downward and briefly intertwines with his. What is going on here? After trying the question in several ways, I conclude that the change of pattern has to do with the language of the dream scene in which the ghosts of Richard's victims curse him and bless Richmond. The scene makes memorable theater but its language is stylized and repetitive and the several monologues are all much like each other. Delta, accordingly, finds no greater affinity here for Shakespeare than for "the gang of four." (It should be observed, however, that the last segments of all, where this scene has been left behind, show a sharp turn back toward Shakespeare.)

In our study of 3 *Henry VI*, Craig and I do not try to single out Shakespeare's collaborator from within the aforesaid gang of four, and I stand by our decision. Kite-flying is fine sport but in attribution studies, as in statistical analysis, it is always better to err on the side of caution. The difficulty presented by these four dramatists is extreme. Each of them wrote at least one play that lies close to Shakespeare's early histories and tragedies in Figure 4, above. But the dramatic corpora of Greene, Marlowe, and Peele are so small and diverse that we are not yet able to obtain robust stylistic signatures. We are able to distinguish each of them from Shakespeare, as with Peele in Figure 6, but not to distinguish them confidently from each other in a complex case like that of 3 *Henry VI*. With Kyd, the problem is even more acute. Only *The Spanish Tragedy* can

[70] Hugh Craig and John Burrows, "A Collaboration about a Collaboration: The Authorship of *King Henry VI, Part Three*," in *Collaborative Research in the Digital Humanities: A Volume in Honour of Harold Short*, ed. Marilyn Deegan and Willard McCarty (Farnham, UK: Ashgate, 2012), 27–65.

be brought to bear. Translations like *Cornelia* yield vexed evidence. And, since the case for Kyd's sole authorship of *Arden of Faversham* remains conjectural, that play should not be used as evidence in support of another conjecture. S. Schoenbaum's dicta are not always incontestable, but his injunction against piling conjecture upon conjecture is well founded.

Craig and I may hope to resume the field with furbished arms and new supplies. Others may find a way of overcoming the obstacles I have described. We are content, meanwhile, to have offered evidence of Shakespeare's role and to have helped to restrict the number of other claimants.

V. Finale

I believe I have shown that Brian Vickers misrepresents our work because he knows little of our methods and has done little to repair the deficiency. Yet on his proper ground, he is a scholar whose work I, like many others, have admired for almost forty years. It is for him to decide whether to put this little affray behind him and accept, as I do and others will, that the relationship between his use of *N*-grams, once "optimized," and ours of words is complementary not adversarial, mutually corroborative at most times, mutually interrogative at others. Like Hugh Craig, Arthur Kinney, and many others, we are fellow laborers in a noble vineyard. We have increasingly effective implements and much real work to do. For my part, therefore, these revels now are ended.