Common-words frequencies, Shakespeare's style, and the *Elegy* by W. S.

Hugh Craig
University of Newcastle, New South Wales
Hugh.Craig@newcastle.edu.au


**Figures and Tables**

1. In 1989 Donald W. Foster published a book about the authorship of an *Elegy* for William Peter, printed in 1612 as the work of "W. S." Foster's book reprints the poem and presents the results of a great many stylistic tests designed to test attributions, especially to the best-known "W. S.," William Shakespeare. The conclusions are very cautious. Shakespeare, according to Foster, could not be ruled out as the author. Indeed, he passed all Foster's tests. But that Shakespeare actually wrote the *Funerall Elegy* "is more than I know," Foster says, and in summing up he will go no further in the positive direction than to say that "[t]here is a possibility, perhaps even a strong possibility, that it was written by Shakespeare" (*Elegy* 7). There the matter rested for some years. At conferences in 1995, and in publications in 1996, however, Foster was emboldened by some new tests, and by a collaborator, Richard Abrams (Thomas Huxley to his Charles Darwin, perhaps), to make an unequivocal claim of Shakespearean authorship. An extensive analysis of rare words shared between canonical Shakespeare and the poem, as well as a study of common linguistic idiosyncrasies, led him to the conclusion that the *Elegy* "belongs hereafter with Shakespeare's poems and plays . . . because it is formed from textual and linguistic fabric indistinguishable from that of canonical Shakespeare" ("Funeral" 1082).

2. Foster's methods have had success elsewhere. He correctly identified the author of *Primary Colors*, an anonymous novel about Bill Clinton's first Presidential election campaign, and then worked as a forensic expert on the authorship of the Unabomber manifesto. Foster also tells the story of how he managed to identify the authors of three anonymous referees' reports on the *Elegy* book, entirely on internal evidence (Foster, *Author* 27-9, 38-40, 43-4).
3. The attribution of the *Elegy* to Shakespeare has been highly controversial. [1] There is a transatlantic dimension to the controversy. British scholars, generally, have been the most outspoken in dismissal of Foster and Abrams' claims, whereas a number of North American collected works of Shakespeare, including Stephen Greenblatt's Norton one, now include a text of the *Elegy*. There have been some interesting more general contributions to thinking about authorial authenticity. Abrams in particular has responded to those who have dismissed the attribution of the *Elegy* to Shakespeare on the grounds that it doesn't sound like Shakespeare by saying that since the quantitative and technical grounds for the attribution are so firm the concept of the "Shakespearean" will simply have to change to accommodate it ("Breaching" 54). There are also arguments about whether or not this is the way Shakespeare might have written in the elegy genre, and about a deliberate or otherwise plainness of style in the disputed poem; and about the biographical hints in the elegy and its dedication and how they might relate to Shakespeare.

4. There have been a number of suggestions for possible authors other than Shakespeare culminating in a very recent article arguing for John Ford on quantitative grounds, and listing a number of forthcoming studies supporting this attribution with other sorts of evidence (Elliott and Valenza, "Smoking Guns"). The present study was designed and executed before the emergence of this single, widely favoured candidate. It aims to explore the common-words evidence for Shakespeare's authorship without considering any alternative candidate, pursuing the problem as a simple "Shakespeare or not" one, following the work already done by Foster, and the objections raised to it by MacDonald P. Jackson ("Editions") and by Ward E. Y. Elliott and Robert J. Valenza ("Glass Slippers").

5. In what follows some of Foster's tests are reconstructed and re-run with a new set of data. The new runs show a much weaker association between Shakespeare's style and that of the *Elegy*, and show up some damaging inconsistencies in the way Foster conducted his tests. In a separate analysis using proximity scores the *Elegy* diverges from the Shakespeare pattern for some groupings of words, where known Shakespeare poems, tested independently, follow the pattern throughout. Together these results cast doubt on the attribution and suggest that Shakespeare is not the author of the *Elegy*. These conclusions stem from a different approach from the most recent Elliott and Valenza study ("Smoking Guns"), in that the question posed concerns a likeness or otherwise to Shakespeare's style *tout court*, and to all other previous contribution to the debate in the sense that the question is pursued here on the basis of a single, broadly based quantitative model of style, using very common words. The study offers a new line of evidence on the question of Shakespeare's authorship, which has, of course, been the main focus of scholarly and other interest in the *Elegy*.
6. The *Elegy* problem presents computational stylistics with some interesting possibilities. At something over four thousand words, the poem is long enough for the kind of work with very common words which has been increasingly practised with computer assistance. (For Foster's book, all the counting was done by hand.) Quantitative internal evidence is the main basis for the attribution: readers' impressions of the *Elegy*’s style have been strongly against Shakespearean authorship. The external evidence is weak - in fact, hardly goes beyond the initials on the published book, since there is no known association between William Peter and William Shakespeare.

7. There are some aspects of the problem which pose difficulties, however. The *Elegy* is non-dramatic verse, and Shakespeare's main output is in drama. The disputed text is itself in a sub-genre (elegy) in which he has written no well-attested work to serve as a comparison. It can be precisely dated -- William Peter was killed, and the *Elegy* was published, in 1612. If is by Shakespeare, it is a very late work, but the Shakespeare poems that survive were either published in the 1590s or seem to have been written no later than the early 1600s.

8. Another difficulty relates to the nature of the Shakespeare canon. Either because his works have unusually high internal variation, or because they show few stylistic quirks, they do not in my experience form strong clusters when tested with established multivariate methods. In general, the Shakespeare texts tend to have middling scores in the main dimensions, making a clear-cut association or disassociation with or from a disputed text hard to achieve. Some of the common approaches of computational stylistics are therefore ruled out. Whereas with some authorial sets high-order principal components in a Principal Components Analysis offer clear authorial separations (Lucy Hutchinson is a case in point [Burrows and Craig]), the Shakespeare set within the current one tends to scatter along the components. Techniques such as Discriminant Analysis tend to give unstable results because there are not enough Shakespeare poetry texts to provide a good training set.

9. Any method to be applied to the problem must work for non-dramatic verse, which is what the target text is; but it is unthinkable that such a method could rest on the poems alone. To match the perceptions of those who are sure they know what "Shakespearean" is, and indeed to match the work of Foster himself, the plays must also be included.

10. After much experimentation with multi-variate techniques like Principal Component and Discriminant Analysis, which, as already hinted, all tended to give mixed or indeterminate results, a suggestion from John Burrows led me to use a
more straightforward measure of proximity between *Elegy* counts and those of Shakespeare, described in detail below, with the data represented by a variety of samplings so as to preserve something of its intrinsic complexity.

11. Measures of proximity must be relative, and the Shakespeare texts have to be placed against a background set of similar plays and poems. *Table 1* shows the poems in the textbase prepared for the study. [2] The *Elegy*, the two Shakespeare narrative poems, and the *Sonnets*, are present as wholes. Then, seeking to assemble a background set representative of the non-dramatic verse of Shakespeare's time, I decided first to include a sample of at least a 2,000-word block from each poet in the *Norton Anthology* (7th edition, 2000) born between 1560 and 1600. This gives a representation of canonical writers of Shakespeare's own generation. Incidentally, this means some women poets as well (Mary Herbert, Aemilia Lanyer, and Mary Wroth); there is one play by a woman in the drama part of the textbase (Elizabeth Cary's *Tragedy of Mariam*). Then I added five elegies over 2000 words long from the control set of elegies Foster himself developed for comparison with the W.S. poem (Foster, *Elegy* 293-309). I also included Chapman's continuation of *Hero and Leander*, to accompany Marlowe's first part. There are just over 190,000 words here.

12. *Table 2* shows the plays. Here the temporal limits are by date of first performance, going from 1580 to 1619; [3] this takes Shakespeare from age sixteen to three years dead (seven years after his retirement), so could well be thought of as a "long" version of his generation. First there are twenty-five of the twenty-eight Wells and Taylor regard as the core of the Shakespeare canon, the ones where there has been the least challenge to the idea that they are his unaided work (109-34). This leaves out some plays that nevertheless are part of what goes to make up the received notion of the "Shakespearean" -- *Macbeth*, for instance, because of evidence of shared authorship, also *Taming of the Shrew* and *Titus Andronicus*. [4] Then there is a full set of the Jonson plays within the time limits (twelve in all), and most of the Middleton ones (ten total). Beyond that, twenty-eight by others, all well-attributed. They give a good representation to each of the four decades included, and include comedies as well as tragedies of various sorts. The main virtue of this collection is its size. It was not entirely designed for the purpose; rather it contains everything available to me that was not anonymous or dubious and was dated between 1580 and 1619. The total is 1.6 million words.

13. The texts all derive from early printed editions rather than modern edited ones. This has the advantage that they can be identified with a single early witness, but makes for considerable editorial intervention to standardise spelling and expand a confusing variety of contractions. A number of words were tagged in the texts to separate homographs, so that *will* is separated into verb and noun forms, *that* into
3.1-42 Common-words frequencies, Shakespeare's style, and the Elegy by W.S.

conjunctive, relative and demonstrative ones, and so on. I collected frequencies of the 120 most common words in a larger set, including anonymous and likely Shakespeare collaborations, and of some more function words that were not so common, and then some more to make up (for instance) complete sets of forms of the verbs to be, to do, etc, of the pronouns, including self and selves forms, making the whole list up to 219. The word-types counted are listed in Table 3 (discussed more fully below).

14. It is sometimes lamented that there is not enough consolidation in the field of quantitative approaches to attribution (Rudman). It seems prudent, then, to begin by attempting to reproduce some of Foster's results. The textbase available to the present writer permits only the counting of very common words, since only these are standardised in orthography in the texts, and there is no tagging beyond selected common word types. Replication is restricted to these variables, therefore. Even here, however, there are serious difficulties. A good example is the simple question of how often that occurs in a single play, The Tempest. Foster publishes a count of 190 (Elegy Table 1.20). This comes from Martin Spevack's concordance to Shakespeare. My own count is 206. Most of the discrepancy comes about because whereas for the concordance that's is a separate heading (fifteen instances), in my text it becomes that is, and each instance adds one to the that count and one to the is count. That leaves a discrepancy of one, which arises because for some reason Spevack omits the that in The Tempest 5.1.255, "Some few odd lads that you remember not."

15. Further problems arise through differing definitions of what is Shakespearean. Titus and the whole of Macbeth, as already noted, are excluded from my set for fear of contamination with another author's or a collaborator's style; among the poems, A Lover's Complaint and "The Phoenix and the Turtle" are omitted, one because there has been a good deal of doubt over its attribution to Shakespeare (Wells and Taylor 124; Elliott and Valenza 189-95), the other because it is too short (sixty-seven lines). Thus there are fewer items in my Shakespeare groups, and my Shakespeare averages differ a little from those given in Foster.

16. Similar divergences could be followed through in almost every word-variable. Foster's system of counting and my own are both defensible -- he is using calculations already published in a well-respected source, while I am intervening in the text to recover the instances buried in contractions -- but the different approaches make a point-for-point parallelling of results impossible. Nevertheless, it follows from Foster's strong claims in his "Case for William Shakespeare" chapter that the Elegy would resemble Shakespeare even with modified data, provided that such data is internally consistent. Foster reports that he was simply unable to find a Shakespeare test that the Elegy could not pass (147). Indeed, there is no reason to think that the method used here for counting words like that would
somewhat result in any less accurate comparison of the style of the *Elegy* with Shakespeare's than Foster's.

17. The first common-word variable he mentions as a serviceable Shakespeare marker is *most*. He suggests that Shakespeare's use of this word increased steadily as his style developed, to a point where it is distinctively frequent (*Elegy* 109). In a later section, Foster considers more generally which word-types among the 30,000 which appear in a Shakespeare concordance might be useful as markers of his style. His procedure is to begin by examining the word-types whose frequencies have the narrowest range in Shakespeare plays (141). There are just nine, he says, "that never deviate in the plays by more than a third from their respective mean frequencies." These, he says, are *and*, *but*, *by*, *in*, *not*, *so*, *that*, *to*, and *with* (141). *By*, *in*, *to*, and *with* he immediately rejects on the ground (it seems *a priori*) that frequencies, though they are notably constant in Shakespeare's plays, are not "remarkably different" from Shakespeare's contemporaries.

18. Jackson thought that there might have been a degree of "unconscious bias" in this rejection, considering that in Jackson's own analysis the *Elegy* failed tests based on Shakespearean frequencies of the rejected word-variables ("Editions" 260-1),[5] so it seems worthwhile for the present study to keep these words in play.

19. Foster then adds *like* to his list, on the grounds (again, it seems, *a priori*) that "no two writers express comparative language in quite the same fashion" (144). He thus arrives at a set of seven common words whose use should provide suitable tests for the likeness of a work of unknown authorship like the *Elegy* to Shakespeare. These form the basis for a comparison between the patterns of use in the disputed poem and his "cross-sample" of forty poems from 1610-13, illustrated in his Tables 1.18 and 1.19. The *Elegy* passes all the tests, even those on the extended list (including metrical, syntactical and figures-of-speech ones, which the present data will not support), while none of the cross-sample poems performs anything like as well.

20. Before attempting to replicate the common-words tests from this longer list, we may examine Foster's suggestion that his chosen seven variables are reliable markers of Shakespeare's style, in particular that they are distinctly more reliable than the variables he considers but rejects. Here the principles are that a good Shakespeare marker must show as little variation as possible within Shakespeare, and have as wide variation as possible between Shakespeare's works and comparable ones by other authors. The t-test is a convenient measure combining elements of intra-group and inter-group variation to form a composite score which represents the degree to which differences between the mean values of a variable in two defined groups are both consistent and marked. The difference between the
means of the two groups is divided by the sample standard deviation to produce a t-value. This t-value can then be compared to a table of distribution of such values given the number of samples tested, to give a probability that the two groups of observations belong to the same population on the particular variable. Table 4 shows the probability scores from the t-test for eleven variables (the seven pursued by Foster, with the four he rejects), comparing in turn the means for Shakespeare plays as against those for the plays by other authors in the database, Shakespeare poems as against poems by others, and poems and plays together (“all texts”).

Foster's basis for selecting variables is generally restricted to plays, though frequencies in some poems are included in his discussion of like (144). The table shows that the variables chosen by Foster vary widely in the degree to which they are distinctive to Shakespeare: not is the most distinctive, with a 0.02 probability that the Shakespeare and non-Shakespeare plays belong to the same underlying population, and and is the least so, with quite a high probability (0.72 -- the maximum possible is 1.0) that the two groups are really indistinguishable on this variable. The mean probability for the second group of variables is higher than the first, but there is again wide variation. The probability associated with by is the lowest of any tested (0.00), for instance, suggesting that, far from being of little interest as a Shakespeare marker, it may well be the best one in the set. Looking across the table one might conclude that the seven markers selected by Foster are a more satisfactory set as a whole, but there would be no basis for rejecting all the individual members of the second group holus-bolus -- rather there is a continuum, with members of both sets among the best and the worst-performing markers.

21. It is worth reiterating that Table 4, like Tables 6 and 7 below, is calculated on the basis of different texts differently prepared from Foster's. Striking variations in the treatment of that have already been mentioned. One might add that Foster’s counts of most exclude instances of use as a substantive (Elegy 109), as the present ones do not. His counts of like exclude the verb (144, 147). As it happens, these are tagged in the textbase used here, so the counts in the present analysis can do likewise. On the other hand, Foster's count includes forms such as life-like (144, 147), as my counts do not. An ambiguity arises with to: Foster mentions it in a list of prepositions (141), suggesting that infinitive uses might have been excluded in the work on its pattern of variation in Shakespeare plays, but it seems more likely that he depended here on the counts in the Spevack concordance where there is no separation of homographs. The figures used in Tables 4, 6 and 7, where I aim to follow Foster as closely as my textbase allows, therefore include infinitive uses. [6]

22. Foster's method in this series of trials is to establish ranges for Shakespearean use of these variables, to provide a pass-fail test of authorship. A newly discovered Shakespeare text might be expected to fall within these ranges, especially after adjusting them to account for genre (poem rather than play) and date (very close the end of Shakespeare's career in the case of the Elegy). Foster's Table 1.17 (my
Table 5) shows the highest and lowest counts in the dramatic canon, first, then those limits adjusted by a "poetry-drama ratio" established by dividing the frequency per 1000 words for the five poems included by the frequency for the thirty-eight plays. [7] The four last plays (Coriolanus, Cymbeline, The Winter's Tale and The Tempest) are then used for a "control group", offering more precisely targeted limits, and again those limits are shown adjusted for the difference between poetry and drama indicated by the writer's use elsewhere. Foster's table shows that with each of the five variables included the Elegy falls within the limits of lowest and highest of the control set as adjusted, what he labels the "most probable frequency."

23. Table 6 shows the results of repeating this study with the slightly different data provided by my textbase. The top section shows the results for the same variables as in Table 5, which is reprinted from Foster, and can be closely compared with it. The outlines, as one would hope, are the same. Of the plays identified with highest and lowest counts in the set as a whole ("combined works") the high point for and in the Foster table is Titus Andronicus, which is replaced by Henry V in Table 6 (Titus Andronicus is not included in my base set). The same "control group" of late plays is used in the two tables, but a couple of times there are substitutions in Table 6, for reasons such as the different ways of counting that already mentioned. In his Table 1.16 Foster lists the poetry-drama ratios he found in his larger set, and these can be compared with those in the first column of Table 6; they vary in small ways, but, as an examination of the Elegy columns of the two tables shows, this can make the difference between a pass and a fail for the disputed poem in the Shakespeare tests. In the rest of Table 6 the same data is collected for most and like, included by Foster for his Table 1.18, 1.19 and 1.20, and for the four "prepositional" variables he notes are exceptionally consistent in frequency in known Shakespeare tests, but rejects as insufficiently different in frequency from other authors' usage.

24. Whereas the Elegy passes all five of the tests of Foster's Table 1.17 on his chosen measure ("most probable frequency"), in the revised data it fails on not and that. Counts of not in the Elegy are the same (41 instances, giving a count per thousand words of 9.5) but counts in The Tempest and The Winter's Tale are slightly lower in Table 6, and the poetry-drama ratio significantly lower (compare Foster's Table 1.16), so that the Elegy count is well above the top of the "Shakespeare" range as defined and adjusted. The different definitions of that result in consistently different numbers for this variable, and in Table 6 the Elegy count is below the minimum of the "most probable frequency" range.

25. These two failures suggest a degree of arbitrariness in the Foster range tests. The effects he has singled out are not strong enough and consistent enough to survive a sampling of the base data that is different from his but in its own way defensible.
Inevitably there will be mistakes in counting. I believe, for instance, that Foster counted an extra *and* in the *Elegy*, and missed a *but* (compare Tables 5 and 6), [8] and there will no doubt be undiscovered errors in my own counting. Methods must be sufficiently robust to allow for a margin of error of this kind as well as the variations of a more systematic nature, arising from the definition of the Shakespeare canon and of the variables, already discussed. The sense conveyed by Foster that there is an uncanny or unfailing correspondence between patterns evident in known Shakespeare and the *Elegy* does not survive the present re-examination of part of his analysis.

26. Foster goes on to offer seventeen tests in his Table 1.19 with results for the *Elegy* and for his "Cross-Sample" of comparable poems. The *Elegy* passes all seventeen, a remarkable result. Eight of the seventeen are based on the frequencies of common words, the seven already discussed, and in addition frequencies of *like* as a suffix ("death-like," etc.), listed separately. The range of values used for the test is described as "the most probable frequency for a Shakespearean poem written late in his career" (146), evidently the same as the culminating test in Table 1.17, i.e. the highest and lowest values from the four late plays adjusted by a poetry-drama ratio.

27. Some errors have crept into these calculations. If one works out the "most probable" range for each of the variables from Foster's own figures in Table 1.20, three -- *most*, *like* and *like* as a suffix -- fail. [9] The *Elegy* only passes on these tests by changing the basis for success or failure to something like the range for minimum and maximum for the "control group" of late plays, unadjusted for the poetry-drama ratio. On the other hand, if this had been used to test the other five common-words variables, *that* would have failed (Table 1.17). If Foster had kept a consistent testing regime for his Table 1.19 the *Elegy* would have fallen outside the specified range on some of the test variables and so spoiled its perfect record of "Shakespeareaness."

28. There are, then, some unsatisfactory features in Foster's pursuit of his own protocols for his Shakespeare tests. Elliott and Valenza suggest that "[t]hese tests [of Foster's] have not fared well since 1989" ("Glass" 179). If so, then it is with good reason. We might return now to broader conclusions. The pass rate for the *Elegy* in Table 6, out of eleven in each case, is seven for "combined works," six for the "probable range," five for the "control group" and three for the "most probable frequency." Interestingly, compensating for systematic poetry-drama differences by adjustments to the ranges results in a worse performance -- from seven to six for the combined-works range, and from five to three for the control group ranges derived from the late plays.
29. One can put these various pass rates in context by determining how readily poems of the time known to be by other authors achieve it. Foster does this for his own results in his Table 1.19. Table 7 compares the Elegy's performance on the "most probable frequency" test with three Shakespeare poems and four elegies by other poets. Looking at the overall pattern first of all, Venus and Adonis and The Rape of Lucrece score higher than the poems by other authors, and the Sonnets score as high as the highest of them, suggesting that the tests do have some power to discriminate; the Elegy, however, is at the lower end, indicating if anything a lesser resemblance to Shakespeare than many other non-Shakespearean poems. Even if we restrict ourselves to the five variables Foster chooses for his Table 1.17, there is among the present set one other poem, the Heywood Elegy, that performs as well as the Elegy by W. S., with three passes, a score shared by all three known Shakespeare poems.

30. A re-examination of Foster's common-word evidence thus shows that his tests are not satisfactory. They are too sensitive to small variations in base counts, and manipulation of the test conditions is needed to achieve the preternaturally perfect result which aligns the Elegy so closely with Shakespeare's style. This discounts part of Foster's evidence for Shakespeare's authorship of the Elegy, but it cannot contribute directly to the underlying question of whether or not Shakespeare was the author. In an attempt to test this afresh we can turn to the fullest exploitation of the data available in the present set of texts. It is all very well to cast doubt on another investigator's tests, but are better ones possible? The method adopted here is as follows. The 219 variables already described are counted in the 106 texts. A first question is how many of these variables to use. The composition of the list, as already mentioned, is mixed: a large group defined by frequencies in a larger set of Renaissance texts, then other groups completing sets according to grammatical categories, or comprising function words of particular interest which do not appear among the 120 commonest. Some of these are very rare indeed, so that there are more zero counts than actual occurrences. This in itself may be valuable information -- there might, for example, be an author who never uses a particular word reasonably common in the texts of his contemporaries -- but such information has to be balanced against the distortion introduced by the fact that a zero count in a short text (say, a 1,600-word poem like Jonson's The Famous Voyage) means something different from a zero count in a long play (the longest in the present set is Bartholomew Fair, also by Jonson, over 36,000 words long). It seems prudent, then, to find a upper limit for the number of zeros in the set. If the 219 variables are ranked in this order, we find at the top end eighty-one variables with no zero counts -- there is at least one occurrence in each of the texts -- and towards the bottom variables which occur in only a handful of the texts, and in one case in none of them. If we decide, as an arbitrary cut-off, to discard variables with more than half zero counts -- i.e. 54 or more -- we are left with 194 variables. Ours as a true plural, with fifty-one zeros, is included, but royal-plural our, with fifty-five, is
not. The zero-counts hierarchy also gives a convenient overall ranking for the variables, which then gives a top 20, a top 40 and so on to the full list of 194.

31. The counts of these 194 variables [10] are then put on to a common basis by transforming them into z-scores. A z-score is the difference between a count and the overall mean for that variable, divided by the standard deviation for the variable. Raw counts are thus interpreted as deviations from a mean, scaled according to the degree of variation in the variable. This ensures that the large counts for the more frequent variables will not overwhelm those of the less frequent ones, which might nevertheless contain valuable stylistic information. Counts which are much larger or smaller than the mean of a variable with little variation in the set thus become large z-scores; those close to the mean and on a variable which scatters widely in value become small z-scores. The new table of z-scores, 194 variables and 106 texts as before, is then used to calculate the proximity between each text and every other. The Euclidean distance is calculated for each variable and these results are then added up to give an overall distance between texts. [11] Two texts which score very similarly on the variables will have a low total distance, and vice versa.

32. If we take the Elegy as an example, we can obtain a proximity count for each of the other 105 texts. Each count is a measure of the distance between that text and the Elegy, based on the sum of the distances between the two texts for individual variables. The primary question in the present context is whether this ranking represents a closeness or otherwise to known Shakespeare texts. Do Shakespeare texts appear near the top, suggesting an affinity with his style? As it happens, using 194 variables, the Elegy's closest neighbour is Tancred and Gismund, followed by Rape of Lucrece, then The Revenge of Bussy d'Ambois and Cymbeline, and so on down to Two Angry Women of Abingdon and Jonson's The Famous Voyage in 104th and 105th place. If Shakespeare texts are assigned the value 1 and non-Shakespeare ones the value 2, to make a "Shakespeare" variable, the proximity scores which form the basis for the ranking can be correlated with it to obtain a measure of the degree to which Shakespeare texts as a group tend to be higher or lower on the list. In this case the correlation is 0.24, which can then be compared with other texts. Is this the kind of score a known Shakespeare text gains, or more like a non-Shakespearean one? If this is the Elegy's score for 194 variables, what is the score for other variable sets?

33. Figure 1 plots correlation scores for the Elegy and for the other texts, grouped into significant classes, for the first 20 variables and then for progressively larger variable sets up to 194. Looking at the plays data first (plays are represented by dashed lines), it is evident that the Shakespeare plays are more strongly correlated with the Shakespeare group than are the other plays for all the variable groups.
used. The proximities-correlation method seems not to be sensitive to date -- Shakespeare's late plays are hardly more or less well correlated with the Shakespeare set than the middle or early ones. [12]

34. The method, then, is effective in isolating an authorial effect in the data, at least as far as drama goes. The pattern with Shakespeare's three poems, here treated as one, is a little different. With 20- and 40-word sets the poetry (the solid blue line) is markedly unlike the Shakespeare group, and hardly distinguishable from poems by others (light grey solid lines). Then as more words are added it rises quickly in the correlations until at 80 words it is with the Shakespeare plays, and though it falls below them a little at 140 words and subsequently, it remains well clear and above the texts by non-Shakespearean authors, whether they are plays or poems.

35. The method thus has some success at distinguishing Shakespeare poems from poems by others as well. The pattern for the *Elegy*, the only single text in the chart, is mixed. With 20, 40, 60, and 80 words it follows the pattern of the Shakespeare poems closely, rising to a score just below theirs and coinciding with the score for the Shakespeare late plays. Then it follows a quite distinct trajectory, falling almost as sharply as it had risen down to a low point at 160 words where it is below the score for five elegies by non-Shakespearean authors, and has the lowest score of any on the chart. It rises a little after that but remains firmly with the non-Shakespearean poems. Using the 180 and 194 sets it matches the mean of the five non-Shakespearean elegies tested. The correlation score of 0.24 already noted for 194 variables is, it seems, definitely non-Shakespearean. It is the level of likeness to known Shakespeare texts one expects from non-Shakespearean poems, whether elegies or the other, more mixed group of narratives, lyrics, sonnets, treatises and so on. Curiously, for sixty-word and eighty-word groups the *Elegy* follows a distinctive Shakespeare-poems pattern quite closely, providing the basis for an argument that it is indeed Shakespearean in style; but the effect of adding more words is to bring down its resemblance to Shakespeare, while the known Shakespeare poems maintain theirs. At 140 words and subsequently the *Elegy* is behaving much like the non-Shakespearean elegies.

36. **Figure 2** explores the rise and fall of the *Elegy* correlations around the peak at 80 word-variables a little further. It plots the correlations for the same text and text groups but using proximity scores based on individual variables, a sampling from the 61-80 and 81-100 sets. All the verb forms in these two sets have been plotted, save for the mixed verb-noun variable *love* (Table 3 gives the complete list of variables in the order used in the plots). There are eight from the 61-80 group, plotted to the left of the vertical dividing line, and nine from the 81-100 one, to the right of the line. In Figure 1, the *Elegy*'s correlations with Shakespeare climb steeply from the 60-word-variables entry to the eighty-word-variables one, then
fall steeply to the 100-word-variable point (and continue to fall). Figure 2 gives some idea how this comes about. The highest peaks are in the 61-80 group, and the deepest troughs in the 81-100 one. Looking first at correlations over 0.275 -- these are word-variables where the Elegy count is unusually close to the Shakespeare pattern -- we find two in the 61-80 set (hath and being) and only one in the 81-100 one (see). Hath, an archaic form, is favoured by Shakespeare, especially in the poems, and is notably frequent in the Elegy, while it is rare in the late non-Shakespearean plays. Being, again, follows a Shakespeare pattern in the Elegy.

37. Among word-variables where the Elegy has a strong negative correlation with the Shakespeare group, those below -0.2, for instance, there are none in the first group as against three (come, can, and could) in the second. Simply, then, there happen to be more word-variables with un-Shakespearean patterns in the second group, and in the collective analysis their influence progressively overcomes that of the word-variables with Shakespearean patterns in the first group.

38. Come may be thought of a generic marker -- all the plays come closer to the Shakespeare pattern than all the poems, Shakespeare's own included. Here the Elegy correlation follows the generic pattern, and is only distinguished by being a little lower than the other means. It is easy to imagine that characters frequently urge each other to "come" do this or that or simply "come," where these phatic or merely gestural uses are much rarer in poetry. In the present analysis, therefore, this word-variable has little to tell about authorial matters. Can and could are different. Among the groups, there is some mixing of poetry and plays in the correlation levels, while the Elegy is sharply low in correlations on these variable. Looking at percentage counts of can, we find 0.42 for the Elegy, and a mean of 0.14 for the Shakespeare set as a whole. The comparison for could is even more stark -- 0.28 in the Elegy, compared to a mean for Shakespeare of 0.08 (0.07 in the poetry). This would mean an expectation based on the whole set, or on the poetry, of between two and three instances in the Elegy, where in fact there are twelve.

39. The writer of the Elegy makes frequent recourse to can and could, in a way that is unlike the general Shakespearean pattern. This may be one of those elements that contribute to the widespread view that the Elegy is not Shakespearean in style. There are three instances in the following extract:

But since the summe of all that can be said
Can bee but said that Hee was good: which wholy
Includes all excellence can be displaide,
In praise of Vertue and reproach of Folly:
His due deserts, this sentence on him giues,
Hee dy'de in life, yet in his death he liues: (ll. 531-536, 63)
The instances are inconspicuous, but do make for a slight forcing of the effect towards hyperbole. Something similar occurs with *could*, as can be seen in a passage with, again, three instances close together:

```
The person of this modell here set out,
Had all that youth and happy dayes could giue him:
Yet could not all encompasse him about,
Against th'assault of death, who to relieue him
Strooke home but to the fraile and mortall parts,
Of his humanity: but could not touch
His flourishing and faire long-liu'd deserts,
Aboue fates reach, his singlenesse was such. (ll. 487-94, 60)
```

It would be easy to delude oneself, especially when supported by the statistics, into hearing Shakespearean or un-Shakespearean usages; there would always (as in the present instance) be the possibility of similar effects somewhere in the authentic Shakespeare canon, which would show that Shakespeare could write like this on occasion. As we have seen, there are also cases, even in a quite limited sampling, of unusual patterns (such as frequency of *hath*) which associate the *Elegy* with Shakespeare's style. One could make a case, based on Figure 2, that the *Elegy*'s non-Shakespearean frequencies of *can* and *could* are best explained by generic constraints, bearing in mind that the average counts for these variables in the other elegies in the set also diverge from the Shakespearean pattern. Finally there is no choice but to respect the cumulative weight of the statistical evidence, in which similarities and dissimilarities with the patterns of known Shakespeare are balanced against each other in a way that is as principled as possible in terms of the method, and also performs satisfactorily in separating known Shakespeare from known non-Shakespeare. There seems no reason to discard the information provided by the variables lower down the frequency list; and that information contributes to an overall picture of a style quite unlike Shakespeare's.

40. Figures 1 and 2 show mean results for all entries save for the *Elegy*. Strictly speaking, this is not comparing like with like. We would wish to know, for instance, if other individual texts fluctuated widely in values in a way that is smoothed out in the means. Figure 3 uses the same base data as Figure 1 but shows correlation scores for all the individual poems in the set. It is thus possible to see how the *Elegy* pattern fits with those of all the other poems, remembering that all are equally free to align themselves with the Shakespeare group (or in the case of the each of the Shakespeare poems, with the Shakespeare group excluding itself). Though the three Shakespeare poems all appear at the upper end of the plot once a pattern emerges with the sixty and more clearly still the eighty-variable set, they are by no means unchallenged in likeness to the Shakespeare group generally.
Poems like Nashe's ballad *The Choice of Valentines* and Herbert's homiletic address *The Church Porch* behave as much like Shakespeare as Shakespeare's own poems on this basis. The obvious explanation for this is their resemblances to drama in general, remembering that the Shakespeare set is mainly plays, in a ratio of twenty-five to three.

41. Figure 3, where the *Elegy* is seen against a background of other poems which participate in a test of resemblance to other Shakespeare texts, thus offers a sobering estimate of how far the present evidence can take us on the question of whether the poem is by Shakespeare or not. The system is reasonably successful in finding a cross-genre authorial signature in poems known to be by Shakespeare. These poems maintain a broadly similar pattern of likeness to Shakespeare whatever the variable set (with the exception of general confusion with the smallest sets); but the *Elegy*, though it does follow the Shakespeare-poem pattern with the smaller sets, diverges as more and more words are added, so that overall its behaviour is more like the known non-Shakespeare group. The balance is against Shakespeare authorship. But Figure 3 shows the limitation of this sort of evidence. Some non-Shakespeare poems can seem more like Shakespeare than some known Shakespeare ones; there are only three Shakespeare poems available, with fewer chances therefore for known Shakespeare to diverge in the way the *Elegy* does. One returns to the difficulties inherent in the exercise: Shakespeare wrote no elegies (or other elegies) to compare with the disputed one; it is drama that must be the standard of "Shakespearean-ness," yet the disputed text is non-dramatic verse.

42. Foster's book and articles present univocal and positive quantitative evidence for Shakespeare's authorship of the *Elegy*. By contrast, the common-words proximity analysis described above offers mixed and often confusing results, but, I think, a negative verdict overall. It seems on the whole to confirm readers' impressions that the style of the *Elegy* is unlike Shakespeare's. Cumulatively, the proximities place the *Elegy* with the non-Shakespearean poems -- with all acceptable variables included, the disputed poem is given a score much lower than known Shakespeare poems. The *Elegy* behaves like a Shakespeare poem on many of the variable sets, but not on all; and there is no obvious reason why a new Shakespeare text, if that was what it was, would not follow his other poems and his plays in sustaining a likeness to his work in general throughout the variables range.

Notes

1. A number of contributions to the debate are conveniently collected in Barroll.

2. Clearly it would be desirable to make all the machine-readable texts used
here available for others' use. Copyright difficulties make this impossible at present, but if the various permissions can be obtained I plan to place a complete corpus in the Oxford Text Archive.

3. Dates are from Harbage, Schoenbaum and Wagonheim.

4. *A Midsummer Night's Dream, Much Ado about Nothing*, and *As You Like It* were omitted as no prepared texts were available. *Mariam* was edited and tagged by Louisa Connors, who kindly allowed me to use the text for this study.

5. Abrams has replied to this criticism ("Breaching the canon") and Jackson has responded to the reply (letter).

6. Abrams, responding to Jackson, argues that separating prepositional from infinitive uses of *to* makes for a pattern which brings late Shakespeare and the *Elegy* together ("Breaching" 52, 54n), but it is clear that Foster was using both kinds of *to* together for the analysis in his book (Jackson, letter).

7. I.e., the total number of instances across the poems are divided by the total number of words in these texts and then multiplied by 1,000, and similarly for the plays.

8. Foster himself accepts the inevitability of errors, particularly when counting by hand (*Elegy* 238).

9. The ranges by my calculation are 3.04 to 3.91 for *like* (*Elegy* count 2.08); 0.3 to 0.47 for *like* as a suffix (*Elegy* count 0.23); and 1.01 to 1.55 for *most* (*Elegy* count 2.55).

10. For a table of frequency counts of the 194 word-variables used in the analysis below in the 106 texts, with word-totals for each text, go to <http://www.newcastle.edu.au/centre/cllc/index.html> and click on "Appendices to publications." All the results in this section of the paper can be reconstructed from this table.

11. The proximity table was produced using part of the output from the SPSS "Hierarchical Cluster" analysis tool, specifying "Interval Data" and "Squared Euclidean distance" in the "Measure" menu and "Z Scores" in the "Transform Values" menu.

12. There is of course considerable uncertainty in the dating of the plays in
the set, both the Shakespeare ones and the others. For instance, many of the
dates given vary between the 1964 and 1989 editions of the Annals.
However, the broad chronological groupings used here minimise the
influence of small variations in date: of all the plays included, only The
Merry Wives of Windsor changes grouping if the 1964 Annals dates are used
(it goes from the "early" to the "middle" group).

Works Cited


Responses to this piece intended for the Readers' Forum may be sent to the Editor at L.M. Hopkins@shu.ac.uk.

© 2002-, Lisa Hopkins (Editor, EMLS).