INTRODUCTION
Authorship attribution and Elizabethan drama: qualitative versus quantitative methods

BRIAN VICKERS

Contributor Biography: Professor Sir Brian Vickers is a Distinguished Senior Fellow of the School of Advanced Study, London University, a Fellow of the British Academy, an Honorary Fellow of Downing College, Cambridge, and an International Honorary Member of the American Academy of Arts and Sciences. His publications include *Shakespeare, Co-Author. A Historical Study of Five Collaborative Plays* (Oxford, 2002); *"Counterfeiting" Shakespeare. Evidence, Authorship, and John Ford's* Funerall Elegye (Cambridge, 2002); *Shakespeare,* A Lover's Complaint*, and John Davies of Hereford* (Cambridge, 2007); and *The Collected Works of John Ford,* Vols. II and III (Oxford, 2017), containing the six co-authored plays.

I

The study of authorship problems in Elizabethan drama began in the early nineteenth century with Charles Lamb's delineation of the style of John Fletcher and its differences compared to Shakespeare. In 1808, a few years later, Henry Weber attempted to distinguish the work of each dramatist in their co-authored play, *The Two Noble Kinsmen*, a task executed with more success in 1847 by Samuel Hickson, and in an exemplary edition and commentary by Harold Littledale (1885). The other play by these two dramatists, *Henry VIII*, received its first authorship division from James Spedding in 1850, shortly followed by Hickson. Both scholars' work was taken up by the New Shakspere Society, founded in 1874, which built on the work of the pioneers by publishing many essays on authorship attribution.[1]

The approach used by these and by other scholars up to the present day was *qualitative*: it studied the individual qualities of a play that defined it as an *artefact* and differentiated it from other plays. Qualitative approaches derive from the experience of *reading* plays and experiencing them in the theatre. It analyses the component parts of a play, as Aristotle did in the *Poetics* (fifth century BC). The modern scholar's focus can include the dramatist's use of his sources, his preferred plot structures, characterisation, verse forms and their construction or prosody (run-on lines, feminine endings, pause patterns). She studies language from all relevant aspects: grammar, syntax, vocabulary (social registers, contractions and expletives); and rhetoric, both the tropes, such as metaphor, and the schemes (the placing and repetition of words within a verse line or sentence).

All these approaches can be classified as *manual*, treating the literary text as an *artefact*, in which each component can be studied separately, and a wide range of independent variables tested.

[1] For an account of authorship studies from the 1800s to 2000, see Brian Vickers, *Shakespeare, Co-Author. A Historical Study of Five Collaborative Plays* (Oxford, 2002).

In that paragraph I italicised some key words: 'qualitative', 'artefact', 'reading' and 'manual'. In the 1960s a new approach emerged in the work of Alvar Ellegård on the Junius papers (1962) and that by Mosteller and Wallace on the *Federalist Papers* (1964). Both studies used computers to process their data, and both used statistics as a control. To state the obvious, computers cannot read, but they can sort and count with remarkable speed and accuracy. The consequences for authorship attribution study have been massive. We may celebrate what we have gained but must also register what we have lost. If the literary text is no longer an artefact, the discipline of *reading* has no place. All that the researcher needs is a list of the play's words, often divided into two categories, 'lexical' and 'function words'. The division has no concern with the words' *meaning*, only with their scope. Lexical (or, as they are sometimes called, 'content words'), such as nouns, verbs, adverbs and adjectives are referential and open-class. That is, they are used for purposes of reference to an imagined reality, and they vary in form, singular or plural number. Function words, such as prepositions, conjunctions, pronouns are closed-class, never varying, and exist to give utterances grammatical propriety.

Mosteller and Wallace, as early users of computer-derived quantification, found that lexical words were of little value in distinguishing the main authors of the *Federalist* papers, Alexander Hamilton and Jay Madison, since the topics covered in their contributions were all concerned with the proposed American Constitution. The common vocabulary was one difficulty, the other being the variable frequency of these 'contextual words'. As Mosteller and Wallace recorded, 'Words such as *law*, *executive*, *liberty*, *money*, *trade*, *war* and *states* vary greatly in their rate within a paper' (p. 18). Since the whole aim of quantitative lexical studies is to establish the frequency with which words occur, using contextual words would be a great disadvantage. Mosteller and Wallace evaluated words 'for their ability to discriminate and for their consistency of rate'. On these grounds they came to trust '*function* words – the filler words of the language', many of which 'are not much influenced by the context of the writing' (p. 17; authors' emphasis). Not all function words were suitable, however, for certain types, as they put it, 'are potentially *dangerous*. Personal pronouns and auxiliary verbs, especially with respect to mood and tense, are likely to be related to external details, and inference from them is difficult' (p. 39; emphasis added). The term 'dangerous' there seems an odd choice, but they subsequently explained that function words are 'a fertile source of discriminators' while 'context is a source of risk. We need variables that depend on authors and nothing else' (p. 265). This pioneering example of quantitative attribution based itself on such 'high-frequency words' as *by*, *from*, *to* and *upon*.

Mosteller and Wallace chose 'variables that depend on authors', as if authors were unconstrained by the conventions of grammar as the fundamental framework for conveying meaning. They conceded that some 'sorts of more meaningful words' seem relatively free from context, such as *commonly* and *innovation*, but were unsuitable due to their low frequency of use (p. 17). That instance apart, they were unconcerned with meaning. They also failed to discuss why Hamilton, for instance, used *upon* or *enough* so frequently. Was it a conscious choice, or unconscious? More recent users of quantitative methods, including David Hoover, John Burrows and Hugh Craig, have argued that function words are valuable authorship markers since they are 'beyond the author's control'. But this is to ignore the main function of language, to communicate meaning. The choice of function words is determined by a larger set of conscious choices, the language user's intended communication of meaning. In real life, as in works of literature, function words are crucial in specifying the details of an utterance which enable it to be

understood.[2] Prepositional meanings, for instance, define the conditions of place and time in describing an intended journey: 'I shall go by train to Marseille then I shall take the boat to … arriving on Tuesday at about 6 p.m.' If I am trying to tell someone over the phone where they can find a file in my office, I might say 'It's in the grey cupboard, on the top shelf, probably underneath a pile of blue folders'. The prepositions allow me to define place and time, avoiding ambiguity. My choice of these prepositions may or may not be unconscious, but it is determined by the meaning I wish to convey. Quantitative lexical studies have no use for meaning.

Mosteller and Wallace succeeded in ascribing to Madison authorship of the 12 previously unassigned *Federalist* papers. However, their success does not justify the use of function words to establish authorship markers in works of imaginative literature. Novels use a mixture of narration, direct and indirect speech, in which the characters' utterances are represented as being their own choice. Novelists have traditionally attempted to individualise characters by their style, and have often been valued for their ability to do so. There are exceptions, of course, some of which are deliberate choices on the author's part, such as the novels of Ivy Compton-Burnett which are largely made up of undifferentiated conversation. Other novelists are less concerned with individual characterisation, as in much popular fiction. In Elizabethan drama, playwrights used great invention in creating a range of spoken styles. In *Hamlet*, for instance, Shakespeare individualised Hamlet, Claudius, Polonius, Gertrude, Ophelia, Osric and the Gravedigger: no reader could confuse the speaker of their utterances. It follows that the function words used by Hamlet are Shakespeare's choice for him, and that he made quite different choices for Ophelia, or the Gravedigger. Quantitative attribution methods put all the instances of *to* or *upon* into a statistical or software programme and produce frequencies correct to three decimal points, but to imagine that these figures give reliable authorial markers is a species of wish fulfilment. The study of function words in drama can be highly significant but it must be qualitative, considering them in their local context and in terms of the speakers' varying intentions towards other characters. For instance, Charles Barber's classic essay on the second person pronoun in *Richard III* is a model of this kind.[3]

II

David Auerbach's 'Critique of Quantitative Methods in Shakespearean Authorial Attribution' in this issue addresses a number of these issues, highlighting two general problems. The first is 'the opacity' of the statistical methods that have been used, and the doubts as to whether they are suitable for attribution studies. The second central problem is one that has long bothered me, namely the 'poverty of the input data. By restricting such analyses to a handful of primitive signals such as word frequency and word succession, many of these researchers end up coating fundamentally simple (and untenable) findings in a statistical glaze, disguising the *explanation* for the results in precisely regimented charts and tables. A shift in focus from presentation of results to methodological justification is required.' (p. 2)

To penetrate the 'opacity' of these methods requires a reader with statistical expertise, which Auerbach possesses, but also an awareness of alternative approaches. As he notes, quantitative analyses 'disregard collocations, word order, and word function (except inasmuch as function is reflected in the subset of words chosen). The two

[2] See my essay 'The Misuse of Function Words in Authorship Studies' (forthcoming).

[3] See Barber, '"You" and "Thou" in Shakespeare's *Richard III*', *Leeds Studies in English*, n.s.12 (1981): 273-89; reprinted in V. Salmon and E. Burgess, *A Reader in the language of Shakespearean drama* (Amsterdam and Philadelphia, 1987), pp. 163-79.

questions, then, are whether such large-scale analysis on primitive lexical data can produce a valid authorial 'fingerprint,' and if it can, whether the attribution methodologies employed are sufficient to produce such a fingerprint.' (p. 5)

Auerbach's main focus is on a forty-page essay in the *New Oxford Shakespeare Authorship Companion* by Elliott and Greatley-Hirsch,[4] who attribute *Arden of Faversham* to Shakespeare, offering 'no fewer than four quantitative methodologies'.[5] Auerbach first evaluates their use of the 'Delta' method introduced by John Burrows, which only achieved 'an 85 per cent success rate' in 'placing the correct author in the top *five* candidates' (p. 6; author's emphasis). It can be discounted. The second method used by Elliott and Greatley-Hirsch, 'Random Forests', gave error rates 'hovering between the 10 and 20 per cent misclassification range in all of their tests', making it unreliable. Moreover, this method 'repeatedly decides for Shakespeare as the author of the entirety of *Arden*,' which no reputable scholar has maintained (p. 7).

The third method used by Elliott and Greatley-Hirsch, 'Nearest Shrunken Centroid' (derived from a genetic study of multiple cancer types), 'utilizes the same raw data as the previous methods—single word frequency counts', but 'uses geometric distance in order to gauge the "distance" between works' (p. 8). Linguistic and literary scholars will doubt whether a method used to 'isolate a small percentage of genuinely indicative genes' (approximately 2 per cent of the data set), will be appropriate for identifying 'decisive words on a per-author basis', and Auerbach's account confirms such doubts (pp. 8–9). As he points out, by adopting this technique Elliott and Greatley-Hirsch make authorial attributions for the three 'finalists for *Arden*, Shakespeare, Kyd and Marlowe' which are 'based only on a small set of words used with consistent regularity' (p. 9).

At this point the two contrasting methods I have defined, qualitative and quantitative, diverge absolutely. The latter method reduces the language of a complex literary artefact to a set of 'words in a bag', on the assumption that their frequencies of occurrence will provide a reliable authorship identification. The quantitative method abandoned the reading of texts, just as it abandoned considerations of meaning and communication between the characters in the play. When its computations are complete, they are regarded as autonomous and decisive, freeing attribution scholars from the norm of consulting the text. The words are in the bag, the text is a closed or locked book. Quantitative attributionists note that the word 'heaven' appears less frequently in *Arden of Faversham* than in Kyd's acknowledged plays, and that is taken as sufficient evidence to dismiss his authorship claims. The relevant frequencies of occurrence are as follows: [6]

	Spanish Tragedy	*Soliman and Perseda*	*Cornelia*	*Arden of Faversham*
heaven	14	6	31	12
heavens	18	28	33	14

Table 1: Occurrence of 'heaven[s]' in Kyd's plays

[4] Elliott, Jack, and Greatley-Hirsch, Brett, "*Arden of Faversham* and the Print of Many," in Gary Taylor and Gabriel Egan (eds.), *The New Oxford Shakespeare Authorship Companion* (Oxford, 2017), 139-81.

[5] Auerbach omits discussion of their use of Burrows's 'Zeta' method, the deficiencies of which have already been exposed by Pervez Rizvi in 'The Interpretation of Zeta test results', *Digital Scholarship in the Humanities*, 2018, https://doi.org/10.1093/llc/fqy038 [last consulted on 19 December 2018].

[6] I have used the original quarto texts from *EEBO*, the counts were supplied by Rob Watt's programme 'Concordance'.

At first glance it may seem surprising that Kyd's *Spanish Tragedy* and *Cornelia* (a free translation of Robert Garnier's *Cornelie*) should have a balance between singular and plural forms, but despite their obvious differences in subject matter, both are Senecan tragedies set in a Christian context. Kyd's Turkish tragedy has a greater number of plural forms, appropriately enough for a play set in a non-Christian context, yet permitting such formulae as 'Great Soliman, heavens only substitute'. The fact that differentiates *Arden* is that, where Kyd's other tragedies deal with national conflicts, private revenge, and the related issues of crime and punishment, *Arden* is solely concerned with an adultery plot and a murder arising from it. Although the lower incidence of 'heaven' caused Elliott and Greatley-Hirsch to deny Kyd's authorship of *Arden*, it would have been inappropriate for Kyd's bourgeois characters to invoke such concepts as earthly or divine justice. In this play justice is instantly effective, the murderers' blunders allowing the Mayor to catch them and send them to execution. As for the related concept of fortune, seen as an alternative explanation of events, this is used by many characters in *The Spanish Tragedy* (17 instances). In *Soliman and Perseda* the action is interspersed with scenes showing the rivalry between three dominant influences that act as Choruses on the action, each contesting for primacy —Love, Death and Fortune—the personified figure accounts for 33 occurrences of the word, leaving 20 for the characters, slightly more than in *The Spanish Tragedy*. In *Arden* there are only two occurrences, reflecting the great difference between this domestic *crime passionelle* and the Seneca tragedies.

David Auerbach's analysis of the 'Nearest Shrunken Centroid' method reveals alarming conclusions drawn on slender evidence, some of which affect the attribution of whole scenes. Proponents of Shakespeare's authorship ascribe scene 8 to him. Elliott and Greatley-Hirsch also attribute the play to Shakespeare, but their analysis of this scene would contradict their attribution, due to the supposedly low occurrence of three words, 'she', 'sir' and 'mistress'. However, if the quantitative attributionist would deign to open the play text he would find that the scene begins with a soliloquy by Mosby (lines 1-44), followed by a bitter quarrel between Mosby and Alice (45-150). Naturally enough, in this dialogue the most frequent words are the personal pronouns, the more formal 'you' and the familiar 'thou', with their cognates ('your', 'thee', 'thine'). The only two instances of 'she' are spoken by Mosby in his opening soliloquy, referring to Alice in a cold and calculating manner. Since no male of a superior rank is present, 'sir' does not occur; and since no character of a rank inferior to Alice is present, there is no need for 'mistress'. The dramatic and social context determines the choice of words. In any case, as I pointed out earlier, these words are given to characters and cannot be taken as establishing an author's unconscious preferences; they are local, not general, and therefore demand to be treated in qualitative not quantitative terms.

Another misleading attribution based on a single word is 'you', which, Elliott and Greatley-Hirsch claim, is consistently 'over-represented' in Shakespeare and 'under-represented' in Kyd. To begin with, they make a category error by treating 'you' as a single term, when in Elizabethan English it was part of a dyad, 'you' – 'thou'. It has long been recognised that this option, present in all European languages, allows for a standard use of 'you' as a neutral term, connotating respect and inter-personal distance, while 'thou' has both a normal status as a term of closeness and informality and an inverted status as one of anger and insult.[7] In drama it is of great significance when a character shifts from one form to the other, the move from 'you' to 'thou' signalling closer contact, positive or

[7] A large scholarly literature exists on what linguists call the 'T-V' ('Tu' / 'Vous') option.

negative. The relevant statistic, then, is not 'you' on its own, but 'you' in relation to 'thou'. The figures for the 'you' / 'thou' dyad in Kyd's plays are as follows:

	Spanish Tragedy	*Soliman and Perseda*	*Arden of Faversham*
you	167	137	334
thou	160	169	160

Table 2: Occurrence of 'thou'/'you' forms in Kyd's plays

These figures must be interpreted locally, in terms of the interaction between characters.[8] *The Spanish Tragedy* shows an equal balance between distance and closeness, while *Soliman and Perseda* leans more towards closeness, either friendly or conflictual. The preponderance of 'you' in *Arden* signifies that there are more characters in the play who must be addressed with the plural form of respect. Arden himself is addressed as 'you' by his servant Michael and by other characters of a lower social status. Alice Arden, as a respectable wife, is invariably granted the respect form, as is Lord Cheyne in scene 9, and other public figures, such as the Mayor in scene 14. As always in this dyad, the switches from one form to the other between closely related characters, such as the husband and wife, Arden and Alice, or the adulterous lovers, Alice and Mosby, are highly significant of changes in attitude, from intimacy to coldness, or to anger. These changes are part of the play's structured meaning and can only be appreciated locally.

Returning to Auerbach's essay, its findings are extremely critical of the 'Nearest Shrunken Centroid' method. It uses 'highly questionable factors' in its determinations, which are 'likely to be biased', since the amount of data is too small 'to avoid the problem of using suspect indicators' (p. 10). Having been based on a model in genetics, when applied to 'the frequencies of words like "and", "sir", and "you"', it produced an error rate which was 'so much worse than for its original genetic application' as to raise 'the possibility that NSC is the wrong tool' (p. 12). All three methods used by Elliott and Greatley-Hirsch have been weighed in the balance and found wanting.

Their failure brings Auerbach back to the methodological issues with which his paper began. In a section titled 'The Limits of Word Frequency' he notes the 'striking contrast between the apparent statistical sophistication' of the methods used by Elliott and Greatley-Hirsch and 'the ultimately primitive nature of the emergent criteria', restricted to lexical units divorced from the semantic meanings of the words treated. Furthermore, their claim that 'the tests are mutually reinforcing fails as well, as the tests are not independent of one another but in fact are closely related, all tests using the same fundamental word frequency data'(ibid.). As I observed earlier, one of the advantages of qualitative text analysis is its use of analytical methods to isolate separate facets of a literary work, producing independent test results. The final section of David Auerbach's paper, called 'Rebuilding the Foundations', describes the 'recurring pitfalls' of computational authorship methodologies as including 'the disregard of syntactic, semantic, and even lexical characteristics of a text, in favor of pure numerical measurements of base frequency' and the 'lack of differentiation between more and less meaningful authorial markers, weighing all markers by some uniform baseline metric (e.g., frequency)' (p. 14). His negative conclusion is to 'wonder whether quantitative lexical analyses can ever gain the level of certainty required' (p. 15). He balances this verdict with some constructive recommendations for achieving more reliable authorship

[8] See my essay, 'Personal pronouns and relationships in *Arden of Faversham*' (forthcoming).

attributions, and it is fitting that they include basic features of qualitative methods, including 'grammatical construction order and formulation', 'word adjacencies / clusters', and 'bigrams, trigrams, et al.' (p. 15).

III

When new methods are introduced we often hear that they have 'displaced' older ones, which are now rendered obsolete. This may be a successful marketing ploy in the world of consumer goods, but it does not apply to scholarship, where established methods will continue to be used as long as they have value, or until the newer approach excels them in every respect. From the arguments presented in the first two parts of this essay, quantitative methods in authorship attribution have serious, probably insuperable deficiencies. Qualitative methods have proven advantages and, moreover, they can benefit from technological innovations without compromising their main advantage, by which they can deploy several independent methods to a problem and evaluate the success of each.

As I said earlier, computers cannot read: but they can recognise separate words, or identical series of letters or characters. This facility was employed in recent times to identify cases of students plagiarising other people's work either in their essay assignments or in exams. This anti-plagiarism software compares two electronic documents and can be set to highlight every instance where both the reference document and the target document share a word-string, from two words (a bigram), to three (a trigram), four (a tetragram) and so on. This technology is used in attribution studies not to expose plagiarism as such, but to identify distances where a dramatist repeats himself. It is—or should be—a well-known phenomenon that all speakers of natural languages have a set of preferred phrases. It has been shown that the brain naturally processes language into 'chunks' of information, typically three to four words long. When large corpora of actual language use, spoken and written, began to be compiled in the 1960s, their availability coincided with the application of software to produce electronic concordances. Corpus Linguistics, a new branch of the discipline, soon discovered that these 'chunks' or word groups occurred far more frequently than had been realised. Evidently the brain practises an economy of effort by grouping words together.[9] The pioneers of this discipline distinguished two types of phrase, collocations that were widely shared (e.g. 'Have a good day') and those that were unusual or individual.

This distinction is operative in many languages in many eras, as can be seen from Elizabethan drama. Some plays contain large amounts of commonplace phrases ('Yes, my Lord', 'God be with ye') alongside individual formulations. We know that this is a standard phenomenon in natural languages, but its relevance to Elizabethan drama is that when we can identify a sufficient amount of individual phraseology in an anonymously published play matching other securely attributed plays, we have a strong clue to its authorship and further researches can be made. Previously it would have been impossible to define individual usage with any certainty, given the quantity of plays that would have to be read and the limitations of the human memory. But now, using 'WCopyfind', for example, we can learn almost instantly that two plays share 300 phrases, an identification which is complete, automatic, and replicable, fulfilling three criteria for the scientific collection of data. We can then check each phrase against an electronic corpus of all the plays performed in the London public theatres up to 1595, say, using

[9] See, for instance, John Mc Hardy Sinclair, *Corpus, Concordamce, Collocation* (Oxford, 1991); Alison Wray, *Formulaic Language and the Lexicon* (Cambridge, 2002); and Brian Vickers, 'Shakespeare and Authorship Studies in the Twenty-First Century', *Shakespeare Quarterly,* 62 (2011):106-42, especially 134-41.

'search and find' technology, and establish that perhaps 20 phrases might be uncommon.[10] The second part if this process is manual and time-consuming, but it gives the researcher the valuable opportunity to see each match in its linguistic context, a visual confirmation that often allows us to recognise related, non-contiguous words (collocations) making up a longer, and therefore less common phrasal unit.

At the beginning of this essay I introduced the two categories of qualitative and quantitative analysis and the subsequent discussion may have given the impression that they are incompatible. This is mostly, but not completely true. The scholar using quantitative methods cannot start reading the play-text, since he lacks the categories by which he could analyse it—characterisation, language choice, phraseology, prosody, rhetoric. But the qualitative scholar can accumulate large quantities of data which will progressively increase the chances of a successful attribution. This is particularly true with the new methods for identifying individual preferences in phraseology, where dozens, or indeed hundreds, of unique matches can be accumulated. Here modern technology revivifies an older tradition, for in the Victorian period some formidable scholars compiled huge lists of matching phrases that helped to identify the individual dramatists in many co-authored plays of the Jacobean and Caroline periods. The work of E. H. C. Oliphant helped to assign the authorship of many plays in the Beaumont and Fletcher Folios of 1647 and 1679, which recent studies using different methods have largely validated.[11] His contemporary and rival, Robert Boyle, claimed to have identified over a thousand characteristic phrases in Massinger's plays. In my recent edition of John Ford's six co-authored plays, by using anti-plagiarism software I was able to confirm (often more precisely) the accepted authorship ascriptions of all except *The Laws of Candy* (1620). The previous attribution was to Ford alone, and a comparison of the play with the rest of Ford's canon indeed yielded some 200 matches of phraseology. However, a parallel comparison with Massinger's canon identified over 700 matching phrases, showing him to have been the leading writer.[12] In this way qualitative approaches can have a significant quantitative dimension.

The essay by Darren Freebury-Jones included here exemplifies the strength of qualitative analysis by using separate tests to challenge the part-ascription to Shakespeare of *Arden of Faversham* which MacDonald Jackson has been claiming since 1963. Freebury-Jones began with verbal parallels. Jackson had found only four rare links between the quarrel scene in *Arden* and Kyd's two other accepted plays, *The Spanish Tragedy* and *Soliman and Perseda*. Using anti-plagiarism software Freebury-Jones found around 20 unique verbal links.[13] With the help of that software he also found 'almost forty verbal matches' between that scene and other scenes in the play, confirming it as a unified composition. In addition to that software Freebury-Jones could also use the remarkable database prepared by Martin Mueller, that pioneer in data-processing for literary texts, and co-creator of the wonderful Chicago Homer interactive database.[14] Mueller has spent years preparing a database of early modern plays in modern spellings marked up to permit the identification of verbal parallels, *Shakespeare His Contemporaries*, which

[10] I have used the free software on http://plagiarism.bloomfieldmedia.com/wordpress/software/wcopyfind/ and https://www.inforapid.de/html/searchreplace.htm [both last consulted on 19 December 2018].
[11] See, e.g., Vickers, *Shakespeare, Co-Author*, pp. 47-75, 346-7, 378-80.
[12] See Brian Vickers (ed.), *The Collected Works of John Ford*, 5 vols. (Oxford, 2012--), 2: 76-134.
[13] Combining that software with the Rizvi database I have increased that figure to over 70. See Vickers, 'Is EEBO/LION suitable for authorship studies?' (forthcoming).
[14] See http://homer.library.northwestern.edu/ [last consulted on 19 December 2018].

ultimately included over 500 plays.[15] The mark-up allowed Mueller to search for n-grams repeated in plays by the same author, and in blog-posts as long ago as 2009 he validated the extended Kyd canon that I had identified by means of anti-plagiarism software, adding *Arden of Faversham*, *King Leir* and *Fair Em*.[16] Mueller published a spreadsheet listing all the 'tetragrams plus' (sequences of four or more consecutive words), in which, as Freebury-Jones shows, *Arden* is closely linked with *Soliman and Perseda*.[17]

The superiority of the qualitative method, I have argued, lies in its ability to test an attribution through several independent analyses. Freebury-Jones gives an exemplary demonstration of this flexibility by reporting on four additional tests, involving Kyd's language and prosody. The first, of his own devising, refutes Jackson's claim that the frequent use of compound adjectives in *Arden* proves Shakespeare's authorship by showing that Kyd is equally fertile in creating such compounds. He then reports on three modern studies of Kyd's prosody. The first, by Marina Tarlinskaja, an exponent of Russian quantitative prosody, analyses the position of pauses within the decasyllabic verse line.

Where classical prosody defined metres by quantity into several distinct types (iambic, trochaic and so forth), the Russian school introduced a new method, classifying the stress on each syllable as either strong or weak. By counting the emphases in every verse line of a play percentage figures can be calculated for each of the 9 positions. Tarlinskaja's stress profile of scene 8 in *Arden of Faversham* showed a deep dip on syllable 6, a feature which she had found both in Kyd and in Shakespeare's early plays. Yet, confusingly enough, as Freebury-Jones points out, she denied Kyd's authorship.

No such confusion is found in the classic study by Philip Timberlake (1931) of the so-called 'feminine ending' in Elizabethan drama, a pentameter with an extra syllable ('To be or not to be, that is the ques'tion'). Timberlake showed that Kyd was the pioneer in using this more flexible verse line, soon followed by Shakespeare. As Freebury-Jones points out, Timberlake's figures for *Arden* show no great difference between the scenes that Jackson would ascribe to Shakespeare and the rest of the play. The last test that Freebury-Jones cites, the study by Ants Oras (1960) of how Elizabethan and Jacobean dramatists placed pauses within the verse line, showed strong similarities between *Arden* and Kyd's other plays. Freebury-Jones has re-calculated the pauses within *Arden*, showing no difference between Jackson's 'Shakespeare' and 'Non-Shakespeare' scenes. As Freebury-Jones concludes, given that Jackson has elsewhere approved both prosodic studies, 'it is regrettable' that he avoided their findings on *Arden of Faversham*; now the identity of *Arden* with Kyd's other plays can no longer be denied (p. 12).

In describing anti-plagiarism software earlier, I mentioned that it was necessary to manually compare every matching phrase in each play pair with a database of plays performed in a given period. A further complication is that the texts that scholars used in these studies were the original quarto editions in old spelling, a feature that created difficulties for some search-engines. These and other disadvantages have recently been overcome by Pervez Rizvi's publication of a modern-spelling corpus of 527 plays performed between 1542 and 1657, for which he has written software programs identifying every instance of verbal repetition, whether n-grams or non-contiguous collocations.[18] The user can now compare any play in the corpus with all the other plays

[15] See http://www.digitalhumanities.org/dhq/vol/8/3/000183/000183.html [last consulted on 19 December 2018].
[16] See my website: http://www.brianvickers.uk/?page_id=1013 [last consulted on 19 December 2018].
[17] Mueller's data is now available at https://docs.google.com/spreadsheets/d/14wNSJMkE6LqDG1gNF3Ing8f8TI_7PlCpXxSjkcnbdEE/edit?usp=sharing [last consulted on 23 December 2018].
[18] See Rizvi's website: http://www.shakespearestext.com/can/ [last consulted on 19 December 2018].

or with any group selected from a time period or an individual author. The resulting data is fully automated, and complete, facilitating many types of study. It is early days yet, but this resource looks likely to revolutionise authorship attribution of the qualitative kind.

One immediate result of using it is to increase the total number of matches found. For example, where Freebury-Jones, using anti-plagiarism software, found twenty matches between the Quarrel scene in *Arden of Faversham* and Kyd's accepted plays, with the help of Rizvi's database I have increased that total to over 70.[19] A further, welcome result is that Rizvi himself has published a series of essays questioning some aspects of current attribution scholarship. In addition to his seminal paper exposing some hitherto errors and confusions in interpreting the 'Zeta' method's (see note 5 above), Rizvi has shown that the elevation of Marlowe to co-author of the *Henry VI* plays by a team of scholars using 'word adjacency networks' is unreliable.[20] In a further essay Rizvi demonstrated serious errors in MacDonald Jackson's use of nine selected words to claim Shakespeare's part-authorship of *Arden of Faversham*.[21] Finally, he published a trenchant critique of a new method introduced by Gary Taylor, called 'microattribution',[22] which was also the subject of critical evaluation by Darren Freebry-Jones and Marcus Dahl.[23]

The appearance of new methods and the emergence of new names, such as David Auerbach, Ros Barber,[24] Marcus Dahl, Darren Freebury-Jones, and Pervez Rizvi, suggests that authorship attribution studies are in excellent health and perhaps on the verge of a major development. In Francis Bacon's wise words, '[t]he art of discovery grows with discovery'.

[19] See Brian Vickers, 'Is EEBO (Lion) suitable for attribution studies?' (forthcoming).
[20] See Pervez Rizvi, Authorship Attribution for Early Modern Plays using Function Word Adjacency Networks: A Critical View', *American Notes and Queries*, 5 December 2018. https://doi.org/10.1080/0895769X.2018.1554473 (last consulted 23 December 2018). This essay evaluates Santiago Segarra, Mark Eisen, Gabriel Egan, and Alejandro Ribeiro, 'Attributing the authorship of the *Henry VI* plays by word adjacency', *Shakespeare Quarterly*, 67 (2016): 232–56.
[21] See Rizvi, 'Small Samples and the Perils of Authorship Attribution for Acts and Scenes', *American Notes and Queries*, 13 November 2018.
[22] See Rizvi, The Problem of Microattribution', *Digital Scholarship in the Humanities*, 12 November 2018 https://doi.org/10.1093/digitalsh/fqy066 [last consulted 23 December 2018].
[23] See Darren Freebury-Jones and Marcus Dahl, 'The Limitations of Microattribution', *Texas Studies in Literature and Language*, 60 (2018): 467-95.
[24] See Ros Barber, 'Big Data, Little Certainty: Marlowe, Shakespeare, and *Henry VI*' (forthcoming), challenging the claims by Craig and Burrows for Marlowe's co-authorship of *Henry VI*, which won this year's Calvin & Rose G Hoffman Prize for the best essay on Marlowe.