Routledge
Taylor & Francis Group

Shakespeare and the 1602 Additions to *The Spanish Tragedy*: a method vindicated

Brian Vickers

School of Advanced Study and University of London

ABSTRACT

This essay is a response to Gabriel Egan's criticisms in the *Year's Work in English Studies* of the author's "Identifying Shakespeare's Additions to *The Spanish Tragedy* (1602): A New(er) Approach", published in *Shakespeare* 8 (2012): 13–43.

A few years ago I published an essay in this journal entitled "Identifying Shakespeare's Additions to *The Spanish Tragedy* (1602): A New(er) Approach". In it I discussed a longstanding problem: who wrote the five Additions made to Kyd's play in the fourth Quarto edition (1602). Two authors have been proposed, William Shakespeare and Ben Jonson, but since the Additions were performed in 1599, this eliminates Jonson, who was paid for his additions in 1601. My essay confirmed the attribution to Shakespeare made by previous scholars, but enlarged it by identifying 116 word-strings or collocations that appear in the 320 lines of these Additions and only elsewhere in Shakespeare's plays and poems. This essay innovated by making use of modern computer software, and was well received.[1] I was somewhat surprised, therefore, to receive from Gabriel Egan an excerpt from his survey of recent work on Shakespeare textual and authorship issues for *The Year's Work in English Studies*, in which he subjected my essay to extremely harsh criticism (see especially 334–38). My surprise was partly due to the fact that Professor Egan is the co-editor of this journal, who had "accepted this article for publication, on the basis of two approving referees' reports arising from double-blind peer review" (334). The referees were of "world-class" standing, as he informed me in subsequent correspondence. Of course, an editor has every right to disagree with his referees' judgements, but Egan attacked my essay with such vehemence as to call in question either their or his professional competence.

Egan's chief contention was that "Vickers's method does not work" (338), and he dismissed other aspects of my argument as "useless" and "false".

My method, to describe it briefly, uses recently designed anti-plagiarism software.[2] In such a program two documents (in electronic text files) are compared automatically, and every matching collocation is underlined. The program can be set to identify each instance of three contiguous words (a "trigram"), the most useful phrase length for this purpose, or longer and less-frequent word-strings. Clicking on the underlined collocation brings up the context, which often reveals additional, non-contiguous matching words. The next step is to check every match against a canon of plays from the years immediately preceding the target text and extending to the putative author's death. Whereas previous searches for "parallel passages" were limited by the researcher's memory or

subjective judgement, the precision with which anti-plagiarism software identifies matching collocations removes all guesswork or bias, searches the whole text in the fraction of a second, and can be replicated. These are essential qualities for a method to be deemed scientific, and experienced scholars in this field have welcomed its application to authorship attribution studies. Professor MacDonald Jackson, for long the leading attribution scholar in this field, praised my "excellent idea of using plagiarism software to search pairs of plays for shared three-word phrases", and conceded that "plagiarism software … has the advantage over my LION searches of being mechanized, objective, and readily replicable" ("New research"; Jackson, personal communication, 18 April 2008). Using this method I was able to confirm my previous attribution of the anonymously published play *The Troublesome Reign of King John* to George Peele, made on linguistic and literary grounds ("*Troublesome Raigne*"), by identifying over 200 matching collocations with Peele's authentic works. The late Charles Forker, who was editing that play for the Revels series, accepted my attribution and reproduced the whole list in an appendix, acknowledging that "Sir Brian isolated these impressive matches in 2009 by means of a computer program that, in my judgement, not only establishes the attribution beyond cavil but that holds rich promise of further such discoveries in the study of anonymous Elizabethan plays" (6–30, 335–56, xiv). Subsequently I published an essay in *Shakespeare Survey*, a peer-reviewed journal, identifying Kyd as the co-author, with Shakespeare, of *Edward III* ("Two Authors").[3] An essay confirming that *Greenes Groatsworth of Wit* was indeed written by Greene, not Chettle, is currently under review.

In his recent attribution studies, MacDonald Jackson has also based his approach primarily on matching collocations, as located in the LION database (a limitation, as Eric Rasmussen has pointed out, given its much smaller size than EEBO). In his journal articles and his recent book claiming Shakespeare's part-authorship of *Arden of Faversham*, Jackson collected data in the form of "phrases and collocations that are found … not more than five times in the drama of the period 1580–1600" (*Determining* 219). My method, in contrast, is based on unique matches, those that occur only in the target text and in one other play.

In all of my work on this topic I have followed two principles, seldom stated but generally observed by attribution scholars, of which Professor Egan seems unaware. First, the process of identifying the author or co-author of an anonymously published play should be limited to the appropriate period of time. It would be pointless to look for collocation matches 40 years earlier or later than the target work.

The second, unstated principle on which I – like many other attribution scholars – have worked, treats early modern drama as a self-contained linguistic corpus. This is not to suggest that it was cut off from politics, religion, or any other aspect of contemporary life, but that its language was differentiated from that spoken or written outside the theatre. The plays were written for as many as six companies competing for audiences, which monitored each other's repertories and were quick to imitate their rivals' innovations. This fierce competition meant that, during the playing season, in order to provide plays six (or seven) days a week, new works were being constantly commissioned, rehearsed, produced, revived and occasionally printed. Dramatists were continuously turning out scripts, up to half of which were co-authored by anything from two to five playwrights. Inevitably, they would repeat themselves, and also echo phrases used by their co-authors, or from other plays in production, as scholars have shown since the 1870s. The formal limitations of the drama, the relatively short length (*c.* 20,000–24,000 words), the medium (primarily blank verse, interspersed with prose and songs), the nature of the iambic pentameter line, which is inimical to some word forms, the need to produce diction that was appropriate to the social classes of the characters represented but could also be understood by a semi-literate audience at first hearing – all these factors restricted the available range of expression and account for the enormous number of echoes that can be traced within a dramatist's work or between his and that of his fellows, up to the closing of the theatres. Studies have documented how frequently individual authors repeat themselves: Marlowe, Peele, Kyd and Shakespeare in this period; later, Dekker, Fletcher, Webster, Massinger, Ford, Shirley (see Vickers, "Identifying Co-Authors"). Although this has yet to be recognised as a standard and

inevitable feature of early modern drama, it does explain why internal studies of these texts, which directly identify matching collocations or word-strings, produce reliable authorship indicators.

Whether or not Egan was aware of these unspoken principles governing the selection of timescale and genre, he disregarded them. He devoted seven dense paragraphs (336–38) to disputing the first 11 of my 116 cited matches, having decided to test them "by searching in EEBO-TCP and LION to see if any of the phrases were simply common in the period and hence are not decisive in ascribing authorship". However, what Egan defines as "the period" is not the specific time span I announced as the limit of my investigation, but anywhere between 1500 and 1700. Furthermore, many of his instances derive from non-dramatic work in a variety of genres, and are thus irrelevant to my specific research area. For example, he rebukes me for not having noticed other instances of the collocation "of it besides" in six texts, for which he only provides the *Short Title Catalogue* numbers (336). Their dates run from 1535, 1562, 1563, 1580, 1585 to 1587. Checking the works cited in chronological sequence, STC 14842 turns out to be *The dialogue bytweene [Pope] Jullius the seconde, genius, and saynt Peter*; STC 4470 is *Two very notable commentaries* on Turkish affairs, translated from the Italian and published in Budapest; STC 5008 is an attack on the Jesuits; STC 3802 is Richard Bristow, *Demaundes to be proponed of catholiques to the heretiques*; STC 4442 is *The sermons of M. John Calvin on Deuteronomie*; STC 3071 is Thomas Bilson, *The true difference betweene Christian subjection and unchristian rebellion*; and STC 3734 is John Bridges, *A defence of the … churche of England*. None of these texts is relevant to the language used in the London theatres between 1580 and 1600.

When Egan does cite plays, he includes texts published many years before and after 1600. In the drama of that period I had found the collocation "hanged up", used in connection with persons, only in the Additions to *The Spanish Tragedy* and in *2 Henry VI* ("Identifying Shakespeare's Additions" 36 (no. 11)). Egan claimed that

> LION finds it in the anonymous play *Nice Wanton* (first performed 1547–53), Lording Barry's *Ram-Alley* (first performed 1608–10), Thomas Dekker's *I Honest Whore* (first performed 1604), Fletcher's *The Spanish Curate* (first performed 1622), Massinger's *Believe As You List* (first performed 1631), William Stevenson's *Gammer Gurton's Needle* (first performed 1552–63), and Lewis Wager's *Life and Repentance of Mary Magdelene* (first performed 1550–66). (338)

Apart from the Dekker play, all these instances either pre-date or post-date the relevant time span for my enquiry. Earlier Egan had concluded one paragraph by saying that "if we extend the date to the end of EEBO-TCP's range (that is, up to 1699) there are 376 hits in 273 records" (336). No doubt there are, but that is utterly irrelevant to my argument. Egan accuses me of having failed to perform "the necessary negative check of excluding phrases that were simply common *in the writing of the period*" (338). The phrase I have italicised shows the false basis on which Egan has performed this destructive critique. I never claimed to have considered the whole of English printed books between 1500 and 1700, because it would be pointless to do so. Our interest must be in the London public theatre between 1580 and 1600. Intent on spearing his prey, Egan seems not to have realised this, but it makes 99% of his criticisms irrelevant. Authorship attribution studies, as practised by MacDonald Jackson and myself, would be impossible on Gabriel Egan's ignoring of the considerations of period and genre.

I happily concede that some of his criticisms are justified, in that I missed four relevant instances of matching collocations in the drama between 1580 and 1600, and also two additional matches in Shakespeare. I should have noted that the phrase "run to the" also occurs in Marlowe's *The Jew of Malta* (1589), the anonymous *Look About You* (1599), and *Julius Caesar* (1599).[4] I missed an instance of "do you heare me sir?" in Henry Porter's *The Two Angry Women of Abingdon* (1599), and one of "Nay blush not" in Haughton's *Englishmen for My Money* (1598). However, these are trivial omissions in a list of 116 matching collocations, many of them with several repetitions – especially given Egan's acknowledgement that "None of this criticism is evidence against Vickers's central claim that the Additions to *The Spanish Tragedy* are by Shakespeare" (338). That being

the case, it would seem that the main goal of his review was to discredit my method. Egan sums up his critique with this damning assessment:

> The unavoidable conclusion is that Vickers's method does not work and the likeliest reason for that is that the database of electronic texts he is searching is complete for Shakespeare but incomplete for other dramatists … In any case, being a database of plays and nothing else, it is useless for doing the necessary negative check of excluding phrases that were simply common in the writing of the period.

I drew on a database created by Marcus Dahl containing over 400 plays and masques performed between 1580 and 1642 ("Identifying Shakespeare's Additions" 35–43). Egan has criticised this database previously, in *YWES* 91 (390–92) for not containing some minor works by Middleton, but as they all date from after 1600 his criticism of it as "incomplete" is irrelevant here. More scathingly, Egan claims that "being a database of plays and nothing else, it is useless for doing the necessary negative check". That limitation is obvious, but in Appendix 1 of my essay Dr Dahl gave a detailed account of the elaborate checking process used to compare the Additions with these 400 plays, in order to identify "rare trigrams". Egan conveniently ignores this document.

In addition to his blanket dismissal of my methods, Egan makes a number of negative comments that deserve to be rebutted. For instance, he claims that I make "an unsubstantiated assertion of some importance: 'the allocation of co-authorship in the Elizabethan and later periods was usually by scenes' (27), and hence one can test individual scenes by his method. In fact we do not know for sure that co-writers divided their workload by scenes as there is very little evidence" (335). Although evidence is lacking for many matters in early modern drama, the few surviving references to the unit of composition in documents specifying pre-performance co-authorship agreements refer to both Acts and scenes (see Vickers, *Shakespeare, Co-Author* 27–34). Empirical analyses of the resulting plays, furthermore, have proved that dramatists divided up their work in this manner, as in "The Booke of Sir Thomas More" and *Titus Andronicus*, for instance (*Shakespeare, Co-Author* 34–43, 67–72, 87–90, 127–29, 131–32, 148–243). The division and allocation of authorship by scenes, on the basis of distinct differences in prosody, diction, characterisation and dramaturgy, has been an accepted scholarly procedure ever since the mid-nineteenth century, when Spedding divided *Henry VIII* between Shakespeare and Fletcher, and Spalding identified the same co-authors in *The Two Noble Kinsmen* (*Shakespeare, Co-Author* 336–41, 407–08). Subsequent analyses of co-authored plays by Shakespeare, Middleton, Fletcher, Massinger, Ford and other dramatists have shown over and over again that the initial division of labour was by scenes.[5] Even when one co-author can be traced revising or "over-writing" a scene originally composed by a colleague, the divisions are still clear-cut, with the scene as the basic unit. This fact is so familiar to anyone who has worked on this phenomenon in the Jacobean and Caroline period that it is somewhat disingenuous of Egan to claim that "very little evidence" exists.

One final instance of Egan's unjustified dismissal of my method concerns the length of word strings and their relative rarity. The provision of large corpora of actual language use, written and spoken, together with the advent of word-processing programs that could produce concordances gave rise some 40 years ago to the new discipline of Corpus Linguistics. We now know that natural languages make great use of collocations, usually called "n-grams", where "n" is any number from 2 upwards. In my essay I had pointed out that although the trigram constitutes the basic unit for collocation matching, "the occurrence of longer consecutive sequences is an even stronger [authorship] indicator, since a run of four words is statistically rarer, one of five is even rarer, and one of six is rarer still", making them more valuable in identifying authorship ("Identifying Shakespeare's Additions" 29). Egan bluntly dismisses this statement as "false" (336), on the basis of tests carried out by David Hoover on a corpus of nineteenth-century novels, plays and poetry. However, both Egan and Hoover fail to realise that attribution methods which work for one genre do not necessarily work for another. The vast expanses of the Victorian novel, unconstrained in vocabulary or any of the other formal limits within which the Elizabethan popular theatre worked, make an inappropriate comparison. It is an undeniable fact that in linguistic texts, spoken or written, repetition of long word

strings is rare, and in this respect Elizabethan drama was much closer to ordinary language than the novels of Henry James. Anyone familiar with modern linguistics will know that, as one authority puts it, "the frequency of an n-gram drops precipitously with its length. Compared with trigrams, tetragrams are quite rare" (Mueller, personal communication, 30 April 2014). One study based on a large corpus of spoken English showed that trigrams are likely to occur three times more often than tetragrams, 30 times more often than pentagrams, and 60 times more often than hexagrams. Similar ratios can be found in the early Greek epic, as Martin Mueller showed in the wonderful "Chicago Homer" database (edited with Ahuvia Kahane) and in his book on the *Iliad* (esp. 21–34 and 135–72). Ian Lancashire's work on recurring collocations in Chaucer's *Canterbury Tales* has reached the same result.

Of the 116 matching collocations found in the 1602 Additions and in Shakespeare's plays and poems, 26 are bigrams, 53 are trigrams, but some of the most revealing matches are of longer word -strings. The four-word collocation "within this houre + that", appears in the 1602 Additions and nowhere else in plays preceding 1602, but it does occur in *The Winter's Tale* (no. 10) (see "Identifying Shakespeare's Additions" 35–43). The 25 tetragrams found only in the Additions and in Shakespeare's canon include "a thing of nothing" (no. 20), "And things called whippes" (no. 34), "torche + we burne day light" (no. 54), and so on. Pentagrams are even rarer in all corpora, yet the Additions have seven only found elsewhere in Shakespeare's canon, including "pale faced Hecate + the Moone" (56), "the moone shine that night" (62), "orchard + alley up and downe" (no. 87), "this same hand + that stabb'd" (no. 114). There is even one hexagram: "I prie through + crevice of a wall" (no. 44).

The suitability of collocation matching for authorship attribution will soon be established beyond dispute. In recent years, Martin Mueller has been constructing a massive database of English drama between 1580 and 1642, Shakespeare His Contemporaries (SHC), to include 600 plays, a corpus that has been marked up in the same way as the Chicago Homer. Mueller recently performed an experiment in what he has dubbed the "forensic philology" of authorship attribution, using the SHC corpus, then consisting of 548 plays, and studying n-grams that recur in only two plays in the corpus ("dislegomena") ("Repeated n-grams"). Mueller estimated that there were about 660,000 of such repetitions, the majority being trigrams, so he focussed on the more manageable set consisting of four words or more, which amount to 273,000. The plays in the SHC corpus produce some 100,000 pairwise combinations, in which "the average and median values hover around 2 … For all pairwise combinations the median value is one, and the average not much above it. In short, it is quite rare for two plays – texts that are typically between 15,000 and 25,000 words long – to share more than one or two of the dislegomena analyzed here." Mueller then found that 22% of the pairwise combinations sharing seven or more dislegomena involve plays by the same author, a discovery that established a solid statistical base for studying dramatists' self-repetition. "If we look more closely at shared dislegomena by same-author play pairs, we discover that on average plays by the same author share five dislegomena, and the median is four. Roughly speaking, plays by the same author are likely to share twice as many dislegomena as plays by different authors." As Mueller summed up the significance of this discovery: "it sets a framework of expectations within which the evidentiary value of shared dislegomena can be evaluated".

Thanks to Professor Mueller I have had access to his table of unique shared tetragrams in the SHC corpus (Mueller, personal communication, 9 January 2014), from which I cite one interesting result. Two plays in Kyd's acknowledged corpus, *The Spanish Tragedy* and *Soliman and Perseda*, share seven unique tetragrams, exceeding Mueller's norm of five such word-strings occurring in plays by the same author. Having long argued that Kyd, not Shakespeare, wrote *Arden of Faversham*, I was pleased to find that it shared 18 dislegomena with *Soliman and Perseda*. As a colleague of mine observed, "On this basis, *Arden of Faversham* is more like Kyd than the traditionally accepted plays" (Freebury-Jones). Once Martin Mueller's pre-programmed SHC corpus becomes generally available it will be recognised as the most powerful tool yet devised for identifying the authorship of anonymously published plays.

Both methods described here are new, and it is hardly surprising that exponents of established attribution techniques should not welcome a rival who questions their methods and results, as professional reputations and research grants might be affected. But any new method deserves a fair hearing so that it can be properly evaluated, not given a crushing dismissal. I am grateful to the editors of this journal for allowing me the opportunity to vindicate my method.

Notes

1. As Douglas Bruster acknowledged in the *New York Times* (Schuessler), my essay was the catalyst for his decision to include the Additions in the forthcoming edition of *Riverside Shakespeare*. See also Bruster.
2. I use WCopyFind, created by Dr Lou Bloomfield of the Physics Department at the University of Virginia.
3. Richard Proudfoot has accepted my argument for Kyd's co-authorship in his forthcoming Arden 3 edition of *Edward III*, although he sees Shakespeare as the reviser rather than the co-author.
4. All play dates are from Wiggins.
5. See Vickers, *Shakespeare, Co-Author* 48–49, 65–66, 82–83, 162–69, 210–29, 255–64, 277–87, 336–80, 386–93, 404–30; Hoy; Lake; Jackson, *Studies in Attribution*; Vickers, "Identifying", where I give detailed authorship divisions of Ford's six co-authored plays.

References

Bloomfield, Lou. "WCopyfind." *The Plagiarism Resource Site*. Bloomfield Media, n.d. Web. 16 Jun 2016. Print.

Bruster, Douglas. "Shakespearean Spellings and Handwriting in the Additional Passages Printed in the 1602 *Spanish Tragedy*." *Notes and Queries* 60 (2013): 420–24. Print.

Egan, Gabriel. "Shakespeare: Editions and Textual Studies." *Year's Work in English Studies* 91 (2012): 328–410. Print.

Egan, Gabriel. "Shakespeare: Editions and Textual Studies." *Year's Work in English Studies* 93 (2014): 295–362. Print.

Forker, Charles R., ed. *The Troublesome Reign of John, King of England*, by George Peele. Manchester and New York: Manchester UP, 2011. Print.

Freebury-Jones, Darren. "N-grams, Authorship and Influence: Restoring Thomas Kyd's Canon." *Scalable Reading*. Northwestern U. 4 Sep 2015. Web. 16 Jun 2016.

Hoy, Cyrus. "The Shares of Fletcher and his Collaborators in the Beaumont and Fletcher Canon." *Shakespeare Bulletin* 8–15 (1956–62). Print.

Jackson, MacDonald P. *Determining the Shakespeare Canon*. Oxford: Oxford UP, 2014. Print.

Jackson, MacDonald P. "New Research on the Dramatic Canon of Thomas Kyd." *Research Opportunities in Medieval and Renaissance Drama* 47 (2008): 107–27. Print.

Jackson, MacDonald P. *Studies in Attribution: Middleton and Shakespeare*. Salzburg: Universität Salzburg, 1979. Print.

Kahane, Ahuvia and Martin Mueller, eds. *The Chicago Homer*. Northwestern University, n.d. Web. 16 Jun 2016.

Lake, David. *The Canon of Thomas Middleton's Plays*. Cambridge: Cambridge UP, 1975. Print.

Lancashire, Ian. "Chaucer's Repetends from the General Prologue of The Canterbury Tales." *The Centre and the Compass: Studies in Medieval Literature in Honor of Professor John Leyerle*. Ed. R.A. Taylor et al. Kalamazoo: Western Michigan U, 1992. 315–65. Print.

Lancashire, Ian. *Forgetful Muses: Reading the Author in the Text*. Toronto: U of Toronto P, 2010. Print.

Mueller, Martin. *The Iliad*. London: Bloomsbury, 2009. Print.

Mueller, Martin. "Repeated n-grams in Shakespeare his Contemporaries (SHC)." *Scalable Reading*. Northwestern U. 3 Sep 2015. Web. 16 Jun 2016.

Mueller, Martin. "Shakespeare His Contemporaries: Collaborative Curation and Exploration of Early Modern Drama in a Digital Environment." *Digital Humanities Quarterly* 8.3 (2014). Web. 16 Jun 2016.

Rasmussen, Eric. Rev. of *Determining the Shakespeare Canon*, by MacDonald P. Jackson. *Shakespeare Quarterly* 62 (2015): 226–29. Print.

Schuessler, Jennifer. "Much Ado About Who: Is It Really Shakespeare?" *New York Times*, 12 Aug 2013. Web. 16 Jun 2016.

Vickers, Brian. "Identifying Co-Authors." *The Collected Works of John Ford*. Ed. Brian Vickers. Vol. 2. Oxford: Oxford UP, forthcoming, 2016. Print.

Vickers, Brian. "Identifying Shakespeare's Additions to *The Spanish Tragedy* (1602): A New(er) Approach." *Shakespeare* 8 (2012): 13–43. Print.

Vickers, Brian. *Shakespeare, Co-Author: A Historical Study of Five Collaborative Plays*. Oxford: Oxford UP, 2002. Print.

Vickers, Brian. "*The Troublesome Raigne*, George Peele, and the Date of *King John*." *Words that Count: Essays on Early Modern Authorship in Honor of MacDonald P. Jackson*. Ed. Brian Boyd. Newark, NJ: U of Delaware P, 2004. 78–116. Print.

Vickers, Brian. "The Two Authors of *Edward III*." *Shakespeare Survey* 67 (2014): 69–84. Print.

Wiggins, Martin. *British Drama 1533–1642: A Catalogue*. 10 vols. Oxford: Oxford UP, 2012–. Print.