The problem of microattribution

Pervez Rizvi
Independent student

Abstract

Microattribution is the name of a method which has recently started to be used in the attribution of parts of early modern plays. The method seeks to make authorship attributions by using samples of writing consisting of less than two hundred words. This article argues that the method should not be used, fundamentally because it flouts the well-founded scientific insistence on the sufficiency of sample sizes. The article considers two recent applications of the method, showing that huge amounts of evidence were overlooked which would have invalidated the conclusions drawn. Moreover, the article demonstrates that the method is biased in favour of authors with large surviving canons, such as Shakespeare, and it cannot therefore be relied upon.

Correspondence:
Pervez Rizvi, 15 Blake Road, Croydon CR0 6UH, England.
E-mail:
pervez.rizvi@gmail.com

Microattribution is the name of a method which has recently started to be used in the attribution of parts of early modern plays. The method seeks to make authorship attributions by using samples of writing formerly considered too small to be safely used. It first appeared in an article by Gary Taylor, which used it to attribute a passage in *Macbeth* to Thomas Middleton (Taylor, 2014). Taylor then co-wrote an article with John V. Nance, which used a variant of the same method to attempt to distinguish imitation from collaboration (Nance and Taylor, 2015). I shall show that the method has not been correctly applied by either article. I shall also argue that, as a matter of principle, even when it is correctly applied, its results should not be relied upon.

Taylor had attributed a passage of only sixty-three words in *Macbeth* to Middleton by using microattribution. I shall consider that article first, not only because of its seminal nature but also because it allows me to make my objections to the premise of the method. The stated aim in Taylor is 'to analyse every word and every possible combination of words' in the passage to be attributed (Taylor, 2014, p. 244). In summary, the method looks for these combinations of words in other plays, focusing on the ones found only in the plays of Shakespeare and Middleton, referring to these as unique matches. It then makes the attribution of the passage to Shakespeare or Middleton by comparing how many unique matches come from each dramatist's plays.

As Taylor acknowledged, sixty-three words is a sample far smaller than had previously been considered the minimum necessary for authorship attribution (Taylor, 2014, p. 244). Its way around that obstacle was to claim that it had transformed the sixty-three words into a larger sample, by combining them in different ways. Now, even a small number of single words can be used to generate a vast number of combinations. Readers who are not familiar with mathematics may be unaware that even as few as ten words can generate more than a thousand combinations of varying lengths, and the number of combinations more than doubles every time we add a new word to the set.[1] By combining the words into different phrases, including phrases consisting of non-contiguous words, Taylor turned the sixty-three words into what appears to be a much larger sample. Many readers' instincts would have told them that this is too good to be true. The scientific method is to make observations and deduce knowledge from them. The knowledge we

Digital Scholarship in the Humanities, Vol. -, No. -, 2018.
doi:10.1093/digitalsh/fqy066

can deduce is inescapably limited by those observations, as interpreted by our theoretical framework. That is why scientists take pains to collect large sets of observations, sometimes conducting experiments that last for years. The process cannot be shortcut, because creating a large sample from a small one by assembling the elements of the small sample in many different combinations merely increases the number of items of data we have: it cannot increase the amount of knowledge that is available for us to discover. If it were otherwise, scientists would be able to take a small number of observations, for example the effect of a drug on just a few patients, and purport to turn it into a larger number of observations by combining them in various ways. We can imagine our unease if, say, a pharmaceutical company announced that they had tested the efficacy of a new drug on just ten patients, rather than a thousand, by exploiting the mechanical process of combining ten sets of readings into a thousand combinations. If that were a sound procedure, it would trivialize the notion of the sufficiency of sample sizes.

The principle that reliable conclusions must be based on a sufficiently large sample of observations is well-known, even to non-scientists, but it is as well to remind ourselves why. It is because conclusions drawn from small samples are unstable, by which I mean that they can be materially changed by a tiny change in the observations on which they are based. Taylor attributes the *Macbeth* passage to Middleton by finding, for example, that in a set of search results, his nine hits exceed Shakespeare's eight (Taylor, 2014, p. 255). Leaving aside the objection that this is not a statistically significant difference, it might take just one change in the data—for example, the attribution or de-attribution of a line in some play from Shakespeare or Middleton, or even the correction of an error in the original texts—to eliminate this minimal difference between them. It is fair to acknowledge that the differences between the two authors are a little higher in some cases, for example, the nine Middleton hits versus the three for Shakespeare that Taylor finds when it restricts its scope only to Jacobean plays. But even there, the difference is barely statistically significant, and only one change in the underlying data would make it even less significant.[2] It is perilous to base attributions on such fine margins. Taylor attempts to justify its violation of this principle by pointing out that we can easily hear the voice of, say, John Donne or Emily Dickinson, even in their short poems (Taylor, 2014, p. 244). That is true. However, writers, such as Donne and Dickinson, who were writing at different times in different cultures, for different purposes, and using the English language in different stages of its development, can hardly be compared to early modern playwrights. The fact that Taylor needed to invent a new attribution method at all to distinguish between Shakespeare and Middleton in *Macbeth* testifies to the difficulty of telling them apart in short passages. Comparisons with other pairs of poets in other contexts, such as the ones Taylor makes, are too facile.

The reader might point out that, as a matter of fact, the sample in Taylor was not small, since the sixty-three words had indeed been assembled into what we must assume were thousands of combinations. However, this is an illusion. A truly large sample consists of many independent observations. Observations created by combining other observations are not independent. An independent observation could be taken only from the source—in this case, the text of *Macbeth*—not manufactured by the mechanical process of combining the same set of words in different ways. The requirement for a large sample must be satisfied in substance, not just in form.

After that theoretical objection to the premise of the method, I shall consider how it was applied in practice. At the end of some preliminary tests, Taylor had concluded that for the method to work we must count types, not tokens.[3] The argument seems to be that, if we are sure that a passage is either by Shakespeare or by Middleton, then the true author must be the one with the highest count of unique types, even if the count of unique tokens indicates the contrary. However, the argument was derived from validation tests on just two samples of sixty-four and sixty-three words, from *King Lear* and *A Mad World My Masters*. Therefore, it can hardly be said that it is an argument applicable generally to early modern plays.

Table 1 *Macbeth*—Matches unique to Shakespeare and one author: top ten

Author	Number of tokens	Number of types
Shakespeare, William	120	114
Shirley, James	59	56
Heywood, Thomas	35	33
Anonymous	28	27
Brome, Richard	25	24
Jonson, Ben	25	24
Massinger, Philip	22	21
Chapman, George	20	19
Middleton, Thomas	23	19
Davenant, Sir William	18	17

Table 2 *Macbeth*—Matches unique to Shakespeare and one author (1603–23): top ten

Author	Number of tokens	Number of types
Shakespeare, William	40	39
Heywood, Thomas	22	20
Middleton, Thomas	21	17
Chapman, George	14	14
Fletcher, John; Massinger, Philip	13	12
Fletcher, John	11	11
Jonson, Ben	11	11
Rowley, William	10	10
Daniel, Samuel	9	9
Goffe, Thomas	9	8

Nevertheless, let us continue, and see how it is applied to the *Macbeth* passage.

Pervez Rizvi has made available a list of all collocation matches between the *Macbeth* passage and a set of more than 500 early modern plays, provided that each collocation is matched either with Shakespeare alone or with at most one other writer.[4] Rizvi's criteria are more restrictive than in *Taylor*: he searched only for collocations contained in ten-word windows, whereas Taylor searched for 'all' collocations. Despite the restriction, Rizvi's list contains as many as 759 entries, compared to the mere 18 that Taylor reports as having found (Taylor, 2014, p. 254–5). Moreover, the list shows that the passage contains collocations shared with many authors, not just Middleton. A reader of Taylor might not realize that Middleton is not in any way unique in sharing collocations with the *Macbeth* passage that no one else shares. Taylor's decision to show only the matches shared uniquely with Middleton is liable to mislead readers into regarding them as being of special significance, when they are just extracts from a much larger set of matches with many different authors. As well as the more fundamental problems with the method, shown in this article, Taylor fell into the error of running a 'one-horse race' with only one candidate, Middleton, considered as an alternative author of the *Macbeth* passage.[5]

It is easy enough to count tokens from the list published by Rizvi and, by using the filtering and duplicate removal options in Excel, we may count the types as well. By doing this, we can summarize the list of the matches by author. Table 1 lists the top ten authors, ordered according to the number of matching types.

We see that Middleton is at Number 9 in the list, far behind Shakespeare. If we adopt the method in Taylor, which is to decide upon the most likely author of the *Macbeth* passage by seeing who has the highest number of unique matches with it, then Shakespeare wins by a long distance over Middleton, contrary to what Taylor found. It is fair to acknowledge that the database of plays used by Rizvi does not contain *The Witch*, so it is likely to understate the Middleton matches. Nevertheless, that database, derived from EEBO-TCP, is a very substantial sample and, as Taylor acknowledges, larger than the LION database it used (Taylor, 2014, p. 257) Moreover, Middleton's totals are so far below those of several other authors, most notably Shakespeare, that it is unlikely that the inclusion of *The Witch* would have made a material difference. Similarly, it is possible that among the matches with anonymous plays, there are more from plays which have not yet been recognized as Middleton's than there are from plays which have not yet been recognized as Shakespeare's. Even so, the gap between Shakespeare and Middleton is so wide that it is impossible to imagine Middleton coming close to Shakespeare even with the addition of some matches from *The Witch* and from anonymous plays. Taylor drew an incorrect conclusion about the authorship

Table 3 *Macbeth*—Matches unique to Shakespeare and one author (1576–1642): top ten

Author	Number of tokens	Number of types
Shakespeare, William	120	114
Shirley, James	58	55
Heywood, Thomas	35	33
Anonymous	27	26
Brome, Richard	25	24
Jonson, Ben	25	24
Massinger, Philip	22	21
Chapman, George	20	19
Middleton, Thomas	23	19
Davenant, Sir William	18	17

Table 4 Top ten canon sizes

Author	Number of words in canon (excluding speech prefixes)
Shakespeare, William	860,234
Shirley, James	567,222
Jonson, Ben	407,773
Heywood, Thomas	397,264
Anonymous	390,924
Brome, Richard	358,870
Fletcher, John	302,528
Middleton, Thomas	293,941
Chapman, George	284,580
Massinger, Philip	258,191

of the passage because it considered a sample of only 18 matches among the (at least) 759 that are now known to exist.

Taylor went further by separately applying two date filters to the eighteen matches, to reduce them only to matches with plays in the 1603–23 and 1576–1642 ranges. It thereby appeared to confirm its conclusion that the *Macbeth* passage is by Middleton and, once again, the conclusion was incorrect for the same reason. Tables 2 and 3 show the results from Table 1 but with the same filters that Taylor applied.

We see that restricting ourselves to matches only in the 1603–23 range causes Middleton to rise to third in the list, but he is still behind Heywood, whom no one suspects of authoring the *Macbeth* passage. As before, the inclusion of *The Witch* and the attribution of one or two anonymous plays to Middleton might well have pushed him up to second, but he would still have been behind Shakespeare.

Table 5 *Mad World*—Matches unique to Middleton and one author: top ten

Author	Number of tokens	Number of types
Shakespeare, William	57	49
Shirley, James	43	37
Heywood, Thomas	41	32
Middleton, Thomas	34	31
Anonymous	25	24
Brome, Richard	26	23
Fletcher, John	21	19
Jonson, Ben	18	17
Chapman, George	16	16
Massinger, Philip	18	15

The reader will have noticed that broadly the same names appear at the top of each table above. What is the reason for this? The answer is apparent when we look at Table 4, which tells us the number of words in the canons of each of the ten most prolific playwrights of the age.[6]

We can now see that the appearance of the authors in Tables 1–3 reflects the size of their canons. The larger an author's canon, the greater the chance that in some place he used one of the types we are searching for. The method used by the article is pre-disposed to find prolific authors like Shakespeare, Heywood, and Shirley to be the most likely candidate authors for almost any passage. To give an example of this, we can repeat our search using the sample of sixty-three words given in Taylor from *A Mad World My Masters*. These search results have also been provided by Rizvi, who finds no fewer than 683 collocation matches that occur only in Middleton's sole-authored plays or those of at most one other author, once again searching in ten-word windows and taking all word combinations into account, including the most common ones, as Taylor's article does. Table 5 gives the top ten authors, according to the number of matching types.

Middleton is a little higher on the list now, but he is still well behind Shakespeare and Shirley. Using the method in Taylor, we have apparently found not

that the *Macbeth* passage was written by Middleton but that the *Mad World* passage was written by Shakespeare. Of course, I do not mean for this attribution to be taken seriously. These results tell us about the soundness of the method, not about the authorship of those passages.

We have seen that when Taylor applied the microattribution method, it used only a small sample of the matches, among the hundreds available. Had it used all the data I have cited above, it would have assigned all three of the passages it tested to Shakespeare, even the one that everyone agrees is by Middleton. I have already cautioned that such samples are too small for the results to be reliable, as the above *faux* attribution of the *Mad World* passage to Shakespeare shows. In the context of plays, the size of a sample we use in our tests needs to be judged by reference to the number of words in the text, not by the number of combinations into which we can assemble those words. I also want to caution against the risk of tailoring a method too finely to make it give the correct answer in the cases we are using to validate it. For example, Taylor had concluded that 'counting tokens produces the wrong result', but this conclusion was drawn only because a validation test failed when it counted tokens but succeeded when it counted types (Taylor, 2014, p. 252). There is no reason to suppose that the conclusion—in this case, that counting types is more reliable for authorship attribution than counting tokens—is generally applicable, and if we assume that it is, we risk prematurely discarding methods of research which might yet yield good results. A researcher who uses just one or even a handful of test results to fine-tune a method intended for general use is taking a big risk, that those results are not representative of the vast body of early modern drama and will cause the method to be wrongly tuned.

To conclude my discussion of Taylor, I have provided an appendix containing a commentary on each of the eighteen matches that it lists for the *Macbeth* passage. The purpose of the commentary is to show that, even if we accept the method as valid, and even if we disregard its failure to consider the vast majority of matches, the matches it does consider do not support the attribution of the passage to Middleton anyway.

It is unnecessary for us to go into the same detail for Nance and Taylor, since it is subject to the same problems. Nevertheless, as Nance and Taylor modified the method in Taylor, we should consider it. The article performs microattribution separately on sets of 173 words, drawn first from *Titus Andronicus*, then from *The Jew of Malta*, and then from other plays. Whereas Taylor had clearly stated that its aim was 'to analyse every word and every possible combination of words', Nance and Taylor makes matters unclear by stating that it will consider 'every sequence and every close collocation' without specifying what it means by 'close' (Nance and Taylor, 2015, p. 34). The passages it uses consist of 173 words each, so the level of closeness can make the difference between considering hundreds of collocations and considering many millions. It further tells us that it considered 'the first and second word of a passage, then the second and third word, then the string of those first three words together'. Now, in a 173-word passage, there are 172 bigrams and 171 trigrams, making a total of 343. The article then tells us that the Excel spreadsheet containing the 'full' list of combinations it considered had 377 rows (Nance and Taylor, 2015, p. 35). This must mean that it considered only thirty-four collocations, which is the difference between 377 and 343 and which is an inexplicably small number of collocations to consider from a 173-word passage. The article does not list the thirty-four collocations nor does it tell us the criteria by which they were chosen from among the hundreds available.

Nance and Taylor then misleads its readers by saying that its Excel spreadsheet for the 173 words from *Titus Andronicus* contains 30,160 cells and contrasting that with the observation that a count of feminine endings in the passage would yield only 42 cells (Nance and Taylor, 2015, p. 35). I do not understand how it arrived at the number 42 but, be that as it may, the misleading point is that the number of cells is 30,160 because each of the 377 combinations of words was searched for in eighty plays, and 377 multiplied by 80 is 30,160. Nance and Taylor is introducing a confusion between the

number of combinations being searched for and the number of search results. It searched for 377 combinations, and these were very far from being different, since each bigram overlaps with at least one other bigram and is contained inside a trigram, so the highness of the number 377 is itself deceptive. Multiplying it by 80 to get 30,160 gives an impressively large answer, but it does not make the sample size any larger: the test is being done on a sample of only 173 words and nothing can alter that limitation. In the extreme case, I could take just one word from one text and compare it with a million passages taken from a variety of other texts. I would thereby obtain 1 million results, which is even more impressively large than the 30,160 that Nance and Taylor had, but I would still be performing authorship attribution on a sample of just one word and, of course, my results would be worthless.

Moving on from the objection above, we see that it is hard to make sense of the results, as described in Nance and Taylor. It says that it 'checked every word sequence and collocation in the passage against all extant plays dated 1576–1594, and identified those that occur in only one other early play' (Nance and Taylor, 2015, p. 35, footnote omitted). The qualifier 'close' has disappeared here, but the article must mean 'close collocation', since, as we saw above, it searched for only thirty-four collocations and therefore could not have considered the thousands more that are not close. It then confuses matters further by telling us that 'of the eighty early playbooks in this set, twenty-two different plays, by at least ten named playwrights, contain at least one unique collocation'. In this sentence it seems likely that by 'collocation' the authors mean what in their previous sentence they had referred to as 'word sequence and collocation', since it would make no sense to report results only for the thirty-four collocations and ignore the results for the 343 word sequences. That this is indeed what they must have meant becomes apparent from the last sentence of their paragraph, which provides the basis for their conclusion: 'Shakespeare's early plays contain eight of the twenty-eight unique parallels to this passage (29%), more than double any other playwright'. Observe in passing that Nance and Taylor defines uniqueness to mean matches found in *Titus* and just one other play, whereas Taylor had defined it to mean matches found in *Macbeth* and in either Shakespeare or Middleton.

Table 6 *Titus Andronicus*—Matches shared with only one play from 1576 to 1594 (excluding *The Jew of Malta*): top ten

Author	Number of tokens	Number of types
Anonymous	244	221
Shakespeare, William	237	210
Lyly, John	134	124
Greene, Robert	91	86
Marlowe, Christopher	85	76
Wilson, Robert	71	60
Peele, George	59	53
Garnier, Robert	47	42
Kyd, Thomas	47	41
Churchyard, Thomas; Garter, Bernard; Goldingham, Henry	32	30

We thus see that Nance and Taylor took a 173-word speech in *Titus Andronicus*, extracted 377 word combinations and collocations from it, searched for each of them in 80 other plays written between 1576 and 1594, and found 28 unique parallels, of which 8 are with Shakespeare plays. It treated this as confirmation that Shakespeare wrote the *Titus* passage. I am not convinced that these results justify the conclusion drawn from them. However, let us disregard my objection and ask instead if the Nance and Taylor results are correct. We shall now see that they are not correct: as with Taylor, the method found only a small proportion of the unique parallels that were there to be found.

Unlike Taylor, the Nance and Taylor article does not state if it counted tokens or types, since it refers only to 'parallels', so let us again count both. Rizvi has provided a list of collocation matches, within ten-word windows, between the *Titus* passage and plays written between 1576 and 1594, including only the matches that occur in just one play other than *Titus.* Even after excluding matches with *The Jew of Malta*, Rizvi's list contains as many as 1,270 tokens, for 1,176 types, vastly greater than the twenty-eight matches that Nance and Taylor reported (Nance and Taylor, 2015, p. 37). Since Nance and Taylor

Table 7 *The Jew of Malta*—Matches shared with only one play from 1576 to 1594 (excluding *Titus Andronicus*): top ten

Author	Number of tokens	Number of types
Anonymous	214	183
Shakespeare, William	190	156
Marlowe, Christopher	84	72
Lyly, John	87	70
Kyd, Thomas	77	60
Greene, Robert	64	52
Wilson, Robert	50	44
Peele, George	46	43
Heywood, Thomas	44	38
Churchyard, Thomas; Garter, Bernard; Goldingham, Henry	27	23

do not specify how 'close' the words in its collocations were, we cannot be sure that its criteria are identical to Rizvi's. Even if they are not identical, the criteria used by Rizvi are reasonable, since collocations in which all words are contained within a ten-word window can reasonably be regarded as 'close'. From what we saw earlier for the *Macbeth* passage, we should expect that the *Titus* passage will have the most matches with plays by the authors who were the most prolific in years 1576–94. That is exactly what we find, in Table 6.

We see again that the highest number of matches occurs with the most prolific authors of the era. This is true even of anonymous plays. To see this, consider that plays by unknown authors are concentrated disproportionately in the period when *Titus Andronicus* and *The Jew of Malta* were written, since plays written after the sixteenth century were usually printed with the author's name on the title page. The anonymous plays whose matches are so prominent in these results are unlikely to be all by the same author. The set of all anonymous plays is bound to include the work of many dramatists. Yet, even though we looked only for unique collocations, rather than ones common in the language of the era, we found them in high numbers in the anonymous plays. As before, we see that unique matches are plentiful, even with works by authors other than the one whose sample of words we are testing.

After the above demonstrations, it will come as no surprise that the same happens with the 173-word passage from *The Jew of Malta* that Nance and Taylor tested. Looking for close collocations found only in this passage and just one other play in the period 1576–94, Rizvi finds 1,072 matching tokens for 905 types, far exceeding the mere 18 that Nance and Taylor reports (Nance and Taylor, 2015, p. 38). Like *Titus*, the highest number of matches are with anonymous plays and plays by the young Shakespeare, not with plays by Marlowe, as shown in Table 7.

We have now seen powerful evidence that the microattribution technique is a way to discover the most prolific authors of the era, not to discover the author of the passage we are testing. It is not necessary to work any further through Nance and Taylor, since the rest of it uses the same unsound method.

We may take an equally brief look at the follow-up article which used an adapted form of the same method on passages from *Cornelia* and *Edward III* (Cooper *et al.*, 2017). At the start of the article, its authors present some rules or principles which they propose to follow (Cooper *et al*, 2017, p. 147):

(1) Do not assume that you know the author of the anonymously published or disputed target text (in this case, passages from *Edward III* that are not widely attributed to Shakespeare).
(2) Use databases and search engines that are publicly available, so that other scholars can replicate your methods and check your reported results.
(3) Find a method that correctly identifies the known author of a sample of dramatic writing comparable to the target text, distinguishing the known author from all other candidate playwrights working in the relevant period.
(4) Apply the same method to the target text of unknown authorship.

We may note at once that the jump from (3) to (4) is an astonishing leap of logic. There is no rational basis for supposing that a method which has given the desired result for one carefully chosen sample of 173 words can be relied on to give the correct result for all 'comparable' texts, or

indeed any of them. Even if this supposition were credible, Cooper *et al*. provides no guidance on how we might decide which texts are 'comparable' and which are not. What is worse, having stated its principles, Cooper et al. violates them on the same page of the article. It tells readers that its method is 'to identify sequences of two to four consecutive words (''n-grams'') or juxtapositions of two or more semantically significant words within ten words of one another (''collocations'')'. However, it fails to tell us what it means by 'semantically significant words', making it impossible for anyone to replicate its results. Yet, even without being able to replicate those results, it seems likely that Cooper et al. overlooked most of the evidence, since it tells us that it found a mere twenty-four parallels to the passage from *Cornelia* that it tested, despite searching in the 10-year period from 1585 to 1594 (Cooper *et al*., 2017, 148). This is on a par with the mere twenty-eight parallels that, as we have already noted, Taylor and Nance found when it tested another 173-word passage, confirming that the same inadequate search method was employed by Cooper et al. as by its predecessor articles.

John V. Nance also used microattribution in a sole-authored article for the New Oxford Shakespeare *Authorship Companion*, citing Nance and Taylor as the authority for his claim that 'a sample size of 173 words is sufficient to identify the correct author of an uncontested verse speech in drama' (Nance, 2017, p. 265). It is puzzling why Nance added the qualifier 'uncontested': since the method has no knowledge of what is or is not contested by scholars, how can such contests affect the reliability of the method's results? Be that as it may, we may note simply that the conclusions drawn in that article cannot be relied upon either, since the authority it cites has been shown above to be incorrect.

I hope the above evidence has shown that the microattribution technique has been performed in an incorrect way, since its published matches are only a handful from the much larger number of matches available to be found; and that, even when it is performed correctly, it cannot be assumed to tell us the author of the passage we are testing. The long-established scientific insistence on the sufficiency of sample sizes is not an unnecessary fetter on our freedom but, rather, a necessary condition for the correctness of our work.

Appendix

Commentary on *Macbeth* Matches Listed in Taylor, 2014

The commentary below relies on evidence from Rizvi's published lists of *n*-gram and collocation matches.[7]

so. —Ay, sir] Although this trigram is shared only with *The Nice Valour*, as the article says, the collocation of these three words is quite common. For example, *The Taming of the Shrew* has 'Ay, sir, so his mother says'. Moreover, there are several other trigrams which *Macbeth* shares with just one other author. There is no reason to regard such things as authorial markers in the absence of other argument or evidence.

Ay, sir, all] As the article says, the trigram is shared only with Middleton, but the collocation of the words is common, being found in *Cymbeline* and *Coriolanus* among Shakespeare plays. As above, there are other trigrams in *Macbeth* shared with the plays of just one author without there being any reason to regard them as authorial markers.

all this is so. But] As the article says, the pentagram is shared only with Shakespeare. However, pentagrams shared with just one other author are common. *Macbeth* has 'show the best of our', which is shared only with *Cynthia's Revenge* by Jonson. There is no reason to regard these as authorial markers absent other reasons.

but why stands] This trigram is also found in *The Lovesick Court*, *Friar Bacon and Friar Bungay*, *Orlando Furioso* and *The Woman in the Moon*, none of which are by Shakespeare or Middleton. It is not an authorial marker for either of them.

stands . . . amazedly] Apart from the example in the article, 'stand(s)' is followed by 'amaze' or its variants within two words in 71 other places, including in several Shakespeare plays, but in many other authors too. It is not an authorial marker for anyone.

amazedly] The word is used by several authors other than Shakespeare, but not by Middleton. Rizvi has provided the list of words used by Shakespeare but never by Middleton, not even in any of his co-authored plays and not even in any of its variant forms. He has also provided the reverse list, i.e. the words used by Middleton in some form but never by Shakespeare in any form.[8] Each list has thousands of words on it, many not particularly rare, which one author never happened to use. Moreover, it is well-known that Shakespeare used several thousand words just once, as did Middleton, so just one usage or non-usage would have removed that difference between them. Given these considerations, we risk serious error if we pluck a handful of such words and treat them as authorial markers without further analysis.

come sisters... **the air**] The bigram 'come sister(s)' is common, being found in 30 other places in early modern plays. To claim uniqueness for the collocation, the article had to go as far as 15 words further in the text, to find the phrase 'the air'. Given any bigram, we can turn it into a unique collocation if we go far enough out and find some other pair of words that happen not to occur in proximity to that bigram in other plays. The example is of no value.

show the best] The trigram is also found in *Cynthia's Revenge*, *Lodovick Sforza*, *The Weeding of the Covent Garden* and *Mariam the Fair Queen of Jewry*. Please see my comments above for 'amazedly'. What is true for single words is true, *a fortiori*, for n-grams: every pair of authors will have many n-grams which only one of them happened to use. It does not follow that they can be treated as authorial markers without further analysis.

show... **delights**] As well as the example the article gives, the collocation is also found in plays by several other authors. The comments above apply.

the best of our] As with 'show the best' above, the tetragram is found in plays by several other authors.

best NEAR delights] The bigram 'best delight(s)' is found in *Thierry and Theodoret*, *The Courageous Turk*, *The Bird in a Cage* and several other plays.

our delights] The bigram is very common, being found in *If It Be Not Good, the Devil Is in It*, *The Lover's Melancholy*, *The Duke's Mistress*, *The Spanish Tragedy*, *The Woman Hater* and many other plays.

charm the] The same comments apply as for 'our delights'.

charm NEAR air] The collocation is found in the plays of several authors, although not Middleton. As above, that is not enough reason to regard it as an authorial marker for Shakespeare.

give... **sound**] The collocation is indeed rare, and not found in Middleton, but it is found in *Technogamia, or The Marriages of the Arts* ('gave no sound') by Barten Holiday. Even if we regard it as a Shakespeare marker, it does not help the article's attempt to show that the *Macbeth* passage is by Middleton.

that this great] The trigram is also found in *The Alexandraean Tragedy*, *The False One*, *Antonius*, *2 The Iron Age*, *2 The Fair Maid of the West* and *The Maid of Honour*.

welcome pay] The collocation is found not just in Shakespeare and Middleton but also in *1 The Iron Age* ('Thy welcome Cousin here I pay').

The witches dance and] Unsurprisingly, this tetragram is rare, given the rarity of dancing witches in the drama of the period. That makes it of limited use as an authorial marker, especially as it is the kind of phrase any dramatist is likely to use in a stage direction if his play involves a dance by witches.

References

Cooper, K., Nance, J. V., and Taylor, G. (2017). *Shakespeare and Who? Aeschylus,* Edward III *and Thomas Kyd, Shakespeare Survey 70*. Cambridge: Cambridge University Press, pp. 146–53.

Jackson, M. P. (2017). One-Horse races: some recent studies. In Taylor, G. and Egan, G. (eds), *The New Oxford Shakespeare Authorship Companion*. Oxford: Oxford University Press, pp. 48–59.

Nance, J. V. (2017). Shakespeare and the Painter's Part. In Taylor, G. and Egan, G. (eds), *The New Oxford Shakespeare Authorship Companion*. Oxford: Oxford University Press, pp. 261–277.

Nance, J. V. and Taylor, G. (2015). *Imitation or Colloboration? Marlowe and the Early Shakespeare*

Canon, Shakespeare Survey 68. Cambridge: Cambridge University Press, pp. 32–47.

Taylor, G. (2014). Empirical Middleton: *Macbeth*, Adaptation, and Microauthorship. *Shakespeare Quarterly*, **65**, 239–272.

Notes

1 From a set of N different words, we may generate exactly $2^N-(N+1)$ phrases of two or more words, even when we treat two phrases as being the same if they contain the same words albeit in a different order. If we insist on the order of the words as well, the number of phrases is vastly greater. For the mathematics, see http://mathworld.wolfram.com/k-Subset.html (accessed 8 September 2018).

2 If we have twelve types and we repeatedly assign each one to either Shakespeare or Middleton by tossing a coin, one man might get nine out of the twelve types a little more often than 5% of the time, just by chance. For the mathematics, see http://www.statisticshowto.com/probability-and-statistics/binomial-theorem/binomial-distribution-formula (accessed 8 September 2018).

3 Taylor, 252–53. A *type* is any collocation of words, while a *token* is an instance of that type in a text. For example, 'tomorrow' is a type and the speech 'Tomorrow and tomorrow and tomorrow . . .' contains three tokens of that type.

4 All lists provided by Rizvi are to be found at http://shakespearestext.com/micro (accessed 8 September 2018). Many of the matches listed by him are of the most trivial kind, which attribution scholars would not usually take note of. Nevertheless, they comply with the requirement in Taylor to 'not discard any linguistic information. . ...to analyze every word and every possible combination of words' (Taylor, 2014, 244).

5 For the one-horse race metaphor and its relevance here, see Jackson, 2017.

6 These counts were made by Rizvi from his database, derived from EEBO-TCP, of more than five hundred early modern plays.

7 See www.shakespearestext.com/can (accessed 8 September 2018).

8 See the files shakespeare-minus-middleton.htm and middleton-minus-shakespeare.htm at Rizvi's micro site footnoted above (accessed 8 September 2018).