In response to unknown peer reviewers

Please find below the points of criticism that caused the rejection of the paper and my response.

p. 1 It is not explained why this particular list of texts is chosen as the 'ground truth' for the experiments. We have many sole-authored well-attributed plays by Lyly, Shakespeare, Dekker, and some of the others that could be used to give greater reliability. In particular, representing Shakespeare by just two comedies is unreasonable.

The chosen test environment did not come out of the blue, but was carefully considered in a series of preliminary tests, which excluded Greene, Peele, Marlowe and some plays by Lyly. Dekker was indeed not considered as he was just 11 years old when the Queen's Men were founded. To gain greater reliability in representing an author with more than two comedies would have worked indeed – in favour of Shakespeare! It has become an acknowledged truth in stylometry tests that a large author corpus creates a bias and is disadvantageous for authors who are only represented by one play.

p. 2 The precise nature of experiments such as "bootstrap consensus tree" needs to be explained. So must notions such as "culling".

It is true, I could have repeated the comprehensive explanations that Eder, Rybicki and Kestemont put down in their 2016 paper "Stylometry with R. A package for computational text analysis, *R. Journal*, 8(1):107-21. But then Hoover's tests from 2004 which established the excellent suitability of a culling value of 70% and Jack Grieve's test of a variety of variables, where mf3c came out as a top choice, would also have to be mentioned. I did not want to give a lesson in the history of recent stylometry. My guess was that reviewers would be familiar with it.

p. 3 The use of Principal Component Analysis here is invalid: the number of dimensions (1000) vastly exceeds the number of data points (25, although for some reason only 14 are shown in Figure 3). In such a sparsely populated data-space, PCA does not give reliable results. As before, the next method (Rolling Delta) needs to be explained.

The number of documents was reduced to 14, as the totally unsuitable files were chucked out to attain a readable diagram. The assumption that a sparsely populated data-space leads to unreliable results, does not quite fit the case. It is well known that even a noisy PCA plot like Fig. 3 can be used to identify clusters of data. Is it possible to get a reference for using this reason to invalidate Fig. 3? As yet I have not found the published reference that the number of dimensions should not exceed the number of data points. PCA is around since 1901, but this must be new. Besides, I followed exactly the steps provided by the R Stylo program suite which means that R Stylo is discredited by the reviewer's point.

p. 6 The Rolling Window principle seems to be misapplied here, since a window of 5000 words (as here) would be over a quarter of the entire play for most of the plays tested here. There is little hope of catching co-authorship with such a large window, since most times the window will sprawl across a change of author. It doesn't help to move the window forward by only 250 words each time, as seems to be done here -- although that is never explained -- since the window is still taking in far too much of the play.

The purpose of the diagram was to illustrate the three lowest delta values for each 250-word segment. Instead of a 5000-word window another window size could have been taken as well with slight variations, but overall with no different result. The criticism that 'a window of 5000 words would be over a quarter of the entire play' has no object. (*Sir Clyomon* is 24078 words long, one quarter of the play would be 6020 words, not 5000). Smaller windows and their attributions can all be seen in Table 1. In fact, 5000 words is the default in R Stylo, and Eder warns explicitly against smaller windows, as they are not reliable. When in 2017 in a SAT conference Ros Barber criticised Hugh Craig and Arthur F. Kinney [eds., *Shakespeare, Computers, and the Mystery of Authorship* (Cambridge University Press, 2009)] for the large Shakespeare share they had given to *The Jew of Malta* their segments (windows) were only 2000 words long. As yet there is no test to find the optimal window size that returns the collaborators and their true shares.

p. 6 Figure 4 should have alerted the investigator to what has gone wrong in the investigation: the likenesses to 'Clyomon and Clamydes' shown by all five plays rise and fall together. Notice that the y-axis runs only from 26 to 34, which exaggerates the differences between the five 'tracks'. In

"Figure-4-remade.png" I have roughly sketched what this graph would show if y began at zero instead of 26. At this scale, it is clear that the five plays are returning likenesses to 'Clyomon and Clamydes' that are not meaningfully distinguishable.

Again, the reviewer obviously had a different notion of what the window procedure entailed. At first there was a huge table that contained the authored plays in column A. From column B to column BZ each play consisted of a row of delta measurements. From each column (i.e. each 250-word segment) the three lowest deltas with the smallest stylistic difference were extracted. All files that did not have at least one low delta were left disregarded. Of course, there is similarity in the remaining plays, after all most of them are comedies. The three lowest deltas in principle qualify for authorship, this is what Burrows stated in his 2002 paper. The overall result stems from the fact that independent methodologies all come to the same conclusion. The chosen y-axis range serves as a magnifying glass. When the diagram was first produced (starting at zero) the curves were indeed indistinguishable. By the way delta differences between plays can never produce the value zero.

Later addition

The reviewer's request to start the graph at zero fails to recognize the nature of z-scores. This becomes clear when the target text is also part of the reference texts. One could assume that the delta difference is indeed zero as the two texts are identical in the rank and frequencies of their words. However, as the z-score for each word is derived from averaged deviations of the chosen corpus, this is not the case.

$search_clyomonclamydes
[1] 17.19818 16.89652 17.12934 17.80338 18.15845 18.50688 18.47647 17.50881
[9] 17.56028 17.22594 16.98084 16.87014 16.95492 17.07979 16.80889 17.71104
[17] 17.04402 16.76702 16.64461 16.13757 16.13577 17.04437 17.62102 17.32552
[25] 16.46958 16.54332 16.64170 16.52163 16.20751 16.85712 16.45563 16.69641
[33] 16.15327 16.41522 16.39391 16.28482 16.71878 16.87016 16.86893 17.33170
[41] 17.50235 17.43275 17.15884 16.85573 16.23479 16.18201 16.56526 16.95524
[49] 16.81950 17.44312 17.54357 18.42772 18.38683 18.38640 18.67564 18.58419
[57] 17.98205 17.81912 16.58099 16.71187 17.03205 17.14007 16.71603 17.16147
[65] 18.06513 17.76464 17.99317 18.16687 18.68524 19.14526 19.07824 18.97306
[73] 19.46826 20.16870 20.30323 20.41932 20.04257 20.01738 20.54754 20.50072
[81] 20.75579
These values yield an average of 17,5976811 and are the basis for measuring the smallest stylistic differences from suitable reference texts.

Again, the reviewer obviously had a different notion of what the window procedure entailed. At first there was a huge table that contained the authored plays in column A. From column B to column BZ each play consisted of a row of delta measurements. From each column (i.e. each 250-word segment) the three lowest deltas with the smallest stylistic difference were extracted. All files that did not have at least one low delta were left disregarded. Of course, there is similarity in the remaining plays, after all most of them are comedies. The three lowest deltas in

principle qualify for authorship, this is what Burrows stated in his 2002 paper. The overall result stems from the fact that independent methodologies all come to the same conclusion. The chosen y-axis range serves as a magnifying glass. When the diagram was first produced (starting at zero) the curves were indeed indistinguishable. By the way delta differences between plays can never produce the value zero.

p. 7 Weirdly, at this point the investigator starts discussing the scores for Shakespeare plays that were not in the original dataset listed on pages 1-2. Where did these scores come from?

I quote my text from page 7: Of the five plays with lowest delta values four belong to Shakespeare and Kyd who must be seen as collaborators. A brief look at matching n-grams confirms this impression. If we add Pervez Rizvi's summary of word pentagrams and leave out the plays that do not fit the period from 1583 to 1599 we gain more evidence of possible authorship.

Sorry, Pervez Rizvi's data base of 537 English Renaissance Plays is well known among scholars. I am very sorry that Rizvi did not turn up in the References. Mea culpa.

p. 8 "Out of 27 entries Shakespeare accounts for 14". But how many were tested? Shakespeare left us far more plays than anyone else and since the reader isn't told which set of plays were tested here it's impossible to know if or how this predominance was adjusted for.

Please see my comment referring to page 7. For the period from 1583 to 1599 there were 27 plays with a high number of tetragram matches (53 down to 20 matches). The actual list goes down to 1 match and is awfully long. I did not count them all. The purpose was to look at the high numbers, and here about half of the plays belong to Shakespeare. The problem may have been that some of the Shakespeare plays were not recognised as Shakespeare plays, as they are not recorded as such with Wikipedia. (see also: Wer schrieb Shakespeare und was schrieb Shakespeare wirklich? Düren: Shaker, 2023, ISBN 978-3-8440-9249-3).

pp. 9-10 New tests are presented with no description of what they consist of, and then the reader gets the startling conclusion that "The overall rating is unique in that Shakespeare's early plays 'Fair Em' and 'Mucedorus' (anon = A) claim 93 % of all attributions". Very few Shakespearians think that the case for Shakespeare writing 'Mucedorus' has been established beyond reasonable doubt, and that 'Fair Em' is his play is even less widely believed.

Yes, this is the main problem. We have our defined cultural knowledge, captured in Wikipedia. Apparently, it is difficult for human beings to adapt to new information. This has become available with the advent of computers and computer programs like R Stylo. Burrows relied in his basic delta studies on 150 words due to the size of poems. Authorship attributions in the past relied very often on a couple of function words, or rare words believed to be typical of a particular author. Delta, applied to plays, can in theory use all words (70% is better as explained before). The result is based on a much larger data population than ever before and is therefore more reliable than results that are based on just a few function words. I described this problem in "Addressing the Need to Revisit Authorship Attributions in English Renaissance Drama," *Anglistik: International Journal of English Studies*, 33.3 (Winter 2022). Another detailed account that makes use of the erroneous attributions of plays to Marlowe, "Methodological observations concerning word rankings and z-score refinements", *Digital Scholarship in the Humanities*, 2023, <https://doi.org/10.1093/llc/fqad079> addresses the same problem. *Fair Em* and *Mucedorus*, analysed with Rolling Delta, can be found at <http://www.shak-stat.engsem.uni-hannover.de/eauthorfairem.html> and <http://www.shak-stat.engsem.uni-hannover.de/eauthormucedorus.html>. It might be of interest that the German scholar Schücking already suggested that "die schöne Emma" might be a play by young Shakespeare.

p. 13 That the Globe was expensive to build might explain the selling of plays to a publisher, but there's no certainty to this. It's just as reasonable to argue that publication of the company's plays would be a useful form of advertisement for the opening of a new theatre. (The cost of building the Globe, about 700 GBP, was probably about one hundred to three hundred times more than what a publisher paid for a play manuscript, so selling a couple of plays would not make a substantial contribution to the cost.)

Yes, true, advertisement for the opening of the Globe is also probable. The decisive point, however, is that the 1599 printing by Creede who, a year later, printed Shakespeare's *Henry V*, refers back to the 1580s and to Shakespeare as likely author/co-author of *Sir Clyomon.*

May I also add that I found the whole review hair-splitting and destructive. My notion of an academic procedure is different. If there are misunderstandings, unclear formulations or different intellectual outlooks, these can be discussed and cleared. In modern times, with computers this does not have to take as long as the change from a geocentric world picture to a heliocentric one.

I think there is enough stylometric evidence that Shakespeare was active as a playwright from the middle of the 1580s.

Reviewer 3:

The work is of poor quality and is not suitable for publication.

It’s a pity that reviewer no 3 did not give any reasons for his or her negative verdict.