Big Data or Not Enough? Zeta Test Reliability and the Attribution of *Henry VI.*

Introduction[1]

"Shakespeare has entered the world of Big Data," editor of the *New Oxford Shakespeare* Gary Taylor is quoted as saying when the announcement of Christopher Marlowe's co-authorship of the three *Henry VI* plays became international news. The idea of Shakespearean "Big Data" is seductive, but stylometric methods that have proved useful in analysing the large well-attributed corpora of novelists like Henry James may not work as effectively with the small datasets we are forced to rely upon for most early modern dramatists (Hoover 2007; Luyckx and Daelemans 2011; Eder 2015). Though high success rates have been claimed for certain stylometric algorithms when it comes to identifying Shakespeare's style in the work underpinning the *New Oxford Shakespeare* attributions of the *Henry VI* trilogy (Segarra, Eisen et al. 2016, 243; Burrows and Craig 2017, 212), it is reasonable to ask whether we have enough data for those stylometric attributions to be reliable.

[1] I am very grateful to Pervez Rizvi, who provided me with the texts and software that allowed me to run my own Zeta tests, helped me set up the test environment, and responded graciously to my questions as testing progressed, often providing additional pieces of code that would help me to answer them. Most crucially, he also suggested a mathematically sound way to analyse Zeta results. Thanks are also due to peer reviewers of this article who helped me to improve it considerably, one of whom (responding to an early incarnation that contained no Zeta tests) pointed out both that everyone knew that genre affected lexical tests like Zeta, and also that it was an unproven assumption.

There are two key methods on which the joint attribution of *Henry VI* to Marlowe and Shakespeare rests: Word Adjacency Networks or WANs (Segarra, Eisen et al. 2016), and the Zeta test (Craig 2009; Craig and Burrows 2012; Burrows and Craig 2017).[2] Though Segarra et al have attempted to defend the newer method, WANs, against a number of criticisms (Segarra, Eisen et al. 2019), further problems have subsequently been exposed. The method breaks its own interpretive rules and seems to be measuring disparity of canon size rather than genuine stylistic differences (Barber 2019). The other key method underlying the attribution, however, is generally considered reliable; Zeta has been widely used in computational stylistics since its introduction in 2007 (Burrows 2007). In his introductory chapter of the *New Oxford Shakespeare Authorship Companion*, Gabriel Egan describes Zeta as 'by some way the most powerful general-purpose authorship tool currently available' (Egan 2017, 45). Yet certain questions about Zeta have not until now been satisfactorily answered.

This article offers some independent testing of Zeta. Following criticism of the existing method of Zeta analysis where, in two-author comparisons, text segments are always attributed to one author or the other, this article introduces a new, statistically-sound method for analysing Zeta results which allows segments to be attributed to neither. The purpose of the tests conducted in this study are as follows:

1. To examine the accuracy of Burrows and Craig's claim that Zeta can tell Shakespeare from Marlowe with 99.9% accuracy.
2. To determine whether the varied dataset sizes present in the original *Henry VI* tests might have an effect on Zeta's reliability

[2] Burrows and Craig also used Delta, Iota and PCA tests, but Zeta was the method most often used, and the method that tended to be given the final say, with Delta used for winnowing, and the other two tests chiefly for confirmation.

3. To examine the extent to which a play's genre (and the overall genre of an author's attributed canon) will influence Zeta attribution.
4. In light of the above, to reconsider what Zeta can really tell us about the authorship of the *Henry VI* plays.

The first part of the study demonstrates that despite the more stringent analysis method, the Shakespeare-Marlowe success rates of Burrows and Craig's 2017 validation tests can be closely replicated. However, the other tests conducted in this study demonstrate that the small canons of most Early Modern dramatists, particularly where they are genre-skewed like Marlowe's, do not provide enough data for Zeta to be reliable.[3]

An overview of Zeta

Zeta is a test that processes words that convey content or meaning, often referred to with the seemingly tautological phrase "lexical words", as opposed to "function words" (prepositions, conjunctions, articles etc.). The underlying theory is that one author will favour different words to another author, and those words can be used in attribution studies as a marker of their individual writing style. To establish which words an author favours, Zeta determines a set of "marker words" for a particular author by comparing word-use in a set of texts attributed to them (the base-set) to the word use of another author or set of authors (the counter-set). Zeta then measures the frequency of those "marker words" in a disputed text.

[3] References to "genre" throughout this paper are to the sub-genres of drama (comedy, tragedy etc). Further subdivisions of form (rhymed or unrhymed verse, prose, song) are ignored, just as they were in the studies of under critique.

There is much variability available to researchers in Zeta's set up. You might, for example, determine the marker words of Author A by comparing the words in their canon (or a subset of their canon) against:

- all other authors of the period X
- selected other authors of the period X
- only authors who write in a specific genre (in period X)
- the entire canon of specific author, Author B
- selected works of specific author, Author B

Period X can be varied. Testing involves breaking the disputed text into segments and discovering how many of these marker words each segment contains. Segment size and the size of the marker word set can also be varied. Thus, we have a plethora of different sub-methods we can choose from to create the set of marker words for Author A, each of which will produce a different set of marker words. This, as we'll see, can significantly influence results.

Improving the analysis of Zeta results

Until now, the results of Zeta testing have always been analysed by the "bisector line" method. Pervez Rizvi explains this method clearly, and also explains why it is inadequate to the task of analysing Zeta results (Rizvi 2019, 403-07). He concludes that bisector line analysis is "a demonstrably unsound procedure" and "too crude to be reliable" (2019, 406). The bisector line method determines authorship of each test segment by its distance from the centroids, the centre point of each data cluster (the base-set and the counter-set). The centroids, being means of their cluster, will be unduly pulled in one direction or another by outliers. Rizvi shows that by using these averages we are analysing complex data with reference to only two data points and losing all the information that the size and shape of the data cluster tells us. This is relevant to the texts under consideration. On Craig's *1 Henry VI* graph for Marlowe, for example, the clusters of both Marlowe and non-Marlowe segments are elongated and overlapping (Craig 2009, 63, fig.3.9). Yet segments of *1 Henry VI* have been allocated to Marlowe despite being only fractionally on the "Marlowe" side of the bisector line.

Since Rizvi published his explanation of what he calls the "bisector line fallacy" it has been clear that a mathematically defensible method of analysing Zeta results must be adopted. Perhaps the most obvious method, suggested to me by Rizvi in private correspondence, is to take standard deviation from the mean of each cluster (i.e. the centroid) into account.[4] This allows us to see whether a segment of the test text falls within the normal distribution curve of either the base author or the counter-set and gives it room (at least in a well-designed test) to do neither.[5] The current bisector method, on the other hand, creates a false binary: it will appear to allocate a segment to Author A or Author B when it may not have a close affinity to either set of texts.

In order to establish the bounds of this improved analysis method, it was essential to determine how many standard deviations from the mean would be appropriate. It might be reasonable, for example, to say that for a test segment to be attributed to the base author, it should fall within two standard deviations of the base centroid (covering 95% of base datapoints) and be outside three standard deviations of the counter (covering 99% of

[4] Pervez Rizvi, e-mail message to author, September 19, 2019.

[5] For clarity, I'm referring here to the bell-shaped curves of normal distribution, a standard statistical concept. Like Hugh Craig's distance-from-centroid method, the new interpretative method rests on the assumption of normal distribution. The improvement is that measurements made with the new method are relative rather than (as with the distance-from-centroid method) absolute. If one is making an assumption of normality (as the centroid method does implicitly) then standard deviation must be taken into account.

these datapoints). For each centroid, two standard deviations need to be calculated, one for the x-axis and one for the y-axis. This effectively creates a 'box' which a datapoint can either fall within or without (see Figure 1).

Testing the standard deviation analysis method

I decided to test this analysis model against the same data that Burrows and Craig used to claim a 99.9% success rate for attribution to Shakespeare, and an 83.1% success rate for attribution to Marlowe, by the bisector line method (Burrows and Craig 2017, 212). Replicating their test, I used non-lemmatised texts and counted types rather than tokens, to create lists of 1000 marker words for both Marlowe and Shakespeare, using Marlowe's complete canon of seven plays and the Burrows and Craig set of eight pre-1600 Shakespeare histories and tragedies.[6,7]

Each tested play was treated as anonymous and subjected to "rolling segmentation", a process by which a text is divided into 2000-word segments which advance (in this case) by 200 words at a time, meaning that the first segment represents words 1-2000, the second 201-2200 etc. This process creates numerous overlapping segments. It also means that the words at the start and the end of the text are significantly under-represented, whereas a single word in the central part of the play will be counted many times (up to ten times, in this case). This aspect of rolling segmentation seems to me inherently problematic, yet its effects on Zeta results, as far as I can tell, have not been

[6] In linguistics, lemmatised texts are marked up so that the lemma – the root form of each word – is counted. *Mouse* and *mice* would both count as *mouse*; *hope, hopes* and *hoping* would all count as *hope*. Non-lemmatised texts (used for this test) are the text as published, meaning the algorithm will take different versions of the same root word as different words. An algorithm can also count either types or tokens. Token-counting involves counting every instance of a word in a segment; type-counting simply means scoring one if there are any instances (no matter how many) and zero if there are none.

[7] The data repository is at Barber, R. (2020). BDNE Zeta Dataset. Goldsmiths Research Online. DOI: 10.25602/GOLD.00028390. Further details about the source texts and their preparation are given in the Appendix.

explored or discussed. This certainly merits further investigation. Having rolling segmentation only for the disputed text also means that its segments on the Zeta graph are bound to fall close to each other, since each successful segment contains (in this case) 90% of the words of the previous segment. This gives the impression that the results for the disputed text are more homogenous than they are. Despite what seem to me obvious problems with this method, it was necessary to replicate it in order to reproduce, as closely as possible, the test conditions of Burrows and Craig.

The texts in the base and counter sets were segmented in a "static" manner (such that a 14,000-word play would generate seven 2000-word segments). Each play being tested was withdrawn from the base set (if present) during its test. Segment counts for each play vary slightly from those of Burrows and Craig because of our using slightly different data sources. But as Craig and Kinney state, "provided the same basic procedures are followed" the overall patterns revealed should be "robust enough to survive the variations that will arise from using different texts" (Craig and Kinney Shakespeare, Computers 2009, xviii). Like Rizvi, I used the modern-spelling texts available from the *Shakespeare His Contemporaries* and *Folger Digital Texts* websites with speech prefixes removed, but not stage directions. The STC numbers of the SHC Marlowe texts are identical to those shown in Burrows and Craig's Table 11.4 (Burrows and Craig 2017, 199).

The results of these tests demonstrate that attribution of a tested segment to the base author if it falls within three standard deviations of the base centroid gives far more satisfactory results than if it falls within two. Using two standard deviations from the base centroid, the success rate for attributing Marlowe's plays to Marlowe is only 34.2% and accuracy is not much better for attributing Shakespeare's plays to Shakespeare (53.7%). Allowing a segment to be attributed to the base author if it was within three standard deviations of the base centroid (while being simultaneously outside three standard deviations of the counter centroid) led to markedly higher accuracy for Zeta. Marlowe was attributed segments of his own plays 72.5% of the time by this method (Table 1), and Shakespeare 91.1% (Table 2). Though this figure is lower than Burrows and Craig's 99.9%, it is still impressively high.

[Table 1]

[Table 2]

The Rizvi method is not only more mathematically sound; it makes the results significantly clearer. For example, using the bisector line method, Burrows and Craig allocated 35.8% of *The Jew of Malta* to Marlowe, leaving one to conclude that the remaining 64.2%, falling on the other side of the bisector line, was theoretically attributed to Shakespeare. By the Rizvi method at three standard deviations, Marlowe is allocated a very similar 35.4% of *The Jew of Malta*, but the segments attributed to Shakespeare constitute only 25.6%, with the remainder of the play falling into no man's land.

Using this more mathematically rigorous method of analysis, *The Jew of Malta,* which Burrows and Craig referred to as a "spectacular failure" in their tests, is not Marlowe's only stylistic anomaly (Burrows and Craig 2017, 210). *Doctor Faustus*, 80.9% Marlowe's by bisector line, is only 24.5% his via the Rizvi method using three standard deviations (Figure 1). Visually, we can see how separate the *Faustus* segments are from most of Marlowe's canon, and the lower figure is in line with this reality. Since both plays contain comic scenes otherwise not present in Marlowe's canon, there may be some "towing" effect towards the Shakespeare cluster that is due to the high level of comedy present in the two *Henry IV* plays.[8] Both anomalies merit further investigation.

[Figure 1]

[8] *Doctor Faustus* is available in two versions, the 1604 'A-text' and the 1616 'B-text'. These tests, like those of Burrows & Craig, use the 1604 A-Text, STC 17429. The A-Text and the longer B-text have a long and controversial history and neither text is free from the possibility of co-authorship. See Keefer, M. H. (2006). "The A and B Texts of Marlowe's "Doctor Faustus" Revisited." The Papers of the Bibliographical Society of America **100**(2): 227-57.

Having established a satisfactory rate of reliability with these particular datasets, the remaining plays of Shakespeare's core canon were tested under identical conditions. The results were encouraging. Across 30 Shakespeare plays, including some recently suggested as bearing signs of other hands, the success rate for attributing Shakespeare plays to Shakespeare averaged 91.6% (Table 3). One caveat here: this success rate only applies to this specific combination of base and counter set. It cannot be taken as a general validation for Zeta, whose success, as I will demonstrate, varies markedly under different conditions. But it seems to confirm that the Rizvi method of Zeta analysis, a more mathematically sound method than the bisector line, is sufficiently reliable that we can confidently utilise it more widely.

[Table 3]

Exploring the limits of Zeta

Testing the full canon reveals that Shakespeare has his own, less dramatic but nevertheless surprising anomalous results. Despite an overall success rate of 91.6% across the thirty Shakespeare plays, *Antony and Cleopatra*'s attribution to Shakespeare is only 67.5%, and *Henry V*'s only 65.3% (Table 3). Although it might seem surprising that a tragedy and history have the lowest attribution accuracy, given that Shakespeare's marker words were generated from a set of his histories and tragedies, Marlowe's marker words were also generated from a set of histories and tragedies. These particular plays have sufficient lexical affinities to Marlowe's to drag a third of each of them into no man's land (Figure 2).

[Figure 2]

In this test, the base and counter sets were relatively balanced in both size (a 1.5 multiple) and genre. But as I will demonstrate, Zeta can dramatically fail when small genre-skewed canons are compared with counter-sets that are mismatched in size, which has been a common feature of Zeta tests on the *Henry VI* plays. Where several authors may have contributed to a text, they should ideally be ruled in or out of contention under the same set of conditions. But in practice, sub-methods to derive the set of marker words are usually varied according to the datasets available. Craig's early work on *1 Henry VI*, attempting to determine the possible contributions of Thomas Nashe, Thomas Kyd and Christopher Marlowe, demonstrates how problematic this can be (Craig 2009, 54-62). It is not only that each author's base dataset is of a different size and nature, which is almost inevitable when dealing with the canons of Early Modern authors, but that they are subjected to non-comparable sub-methods of Zeta. As can be seen in Table 4, in the process of assessing the possible contribution of each author to *1 Henry VI*, variations are made in terms of base-set and counter-set segment size, the method of segmentation of the test play, the proportion of it subjected to testing, and the size and composition of the counter dataset used to generate author marker words.

[Table 4]

Though some rationale is given for most of these decisions, the full effect of such variations on the outcome is neither explored nor accounted for. In every test, there is a significant disparity between base-set and counter-set size, with the latter ranging from 3.7 to 67 times the former's size. It is "not possible … to rule out Nashe", with the lowest counter-to-base multiples of 3.7 and 8.5, as the author of Act 1 (Craig 2009, 55). Yet Kyd, whose counter-to-base multiple is never reduced below 16.6, is dismissed as a possible contributor, with the Zeta results defended as grounds to rule Kyd out of contention despite the fact the "results present a direct contradiction of an authorship hypothesis coming from a respected scholar [Brian Vickers] and supported by a carefully established case" (Craig 2009, 56, 58).

Craig's confidence that the results overturn a "carefully established" case rests on apparent validation for the method presented in an earlier chapter (Craig and Kinney

Methods 2009, 19-24) as well as more specific validation in his own. But Craig's limited testing on three early Shakespeare histories and three early comedies was neither comparable nor sufficient (Craig 2009, 45-7).[9] Not comparable, since it relied on combining a "lexical words test" (presumably Zeta) with a PCA function word test. And not sufficient because only 43 of the 58 2000-word segments were classified as "Shakespeare" by the combined methods, a failure rate of 26 per cent (unless far more of the Shakespeare canon is authored by others than is generally recognised).

In addition, even relative success with a base set of 27 Shakespeare plays against a counter-set of 109 "single-author, well-attributed non-Shakespeare plays" should not lead us to assume that a three-play base set, compared against a disproportionately large counter-set — the conditions under which Kyd was ruled out as a co-author of *1 Henry VI* — is an accurate test environment. Both these tests and the earlier validation tests on *Coriolanus* and *Hengist King of Kent* have a relatively large base-set, and a counter-base multiple of around 3.5. As such, they are essentially validating an entirely non-comparable sub-method.

Craig understands there might be a problem with a large disparity between base and counter dataset sizes, saying "Kyd's two unquestioned plays are really too few for trustworthy results, certainly when compared to a very large assortment of plays by various other authors" (57). While Craig consciously reduces the disparity in size between the two data sets (from a 62-fold difference to a 16-fold difference) by expanding the Kyd plays to include *Soliman and Perseda*, and reducing the counter set to 48 plays, one wonders why he did not reduce it further, as he did with Nashe, re-testing it against Shakespeare pre-1600 dramas (a six-fold difference for three-play-Kyd).

[9] The plays tested were *Richard II, Richard III, King John, The Two Gentlemen of Verona, The Comedy of Errors*, and *Love's Labour's Lost.*

Effect of counter-base size disparity on Zeta results

Independent Zeta validation tests show that a significant counter-base size disparity leads to unpredictable results. In a critique of the Zeta tests conducted on *Arden of Faversham* by Jack Elliott and Brett Greatley-Hirsch, Pervez Rizvi found that *Richard III* does not register as Shakespeare's when using the same method they used to attribute *Arden* (Elliot and Greatley-Hirsch 2017; Rizvi 2019, 416). Their dataset of 35 plays of the period 1580-1594, with *Richard III* withdrawn for testing, provides a Shakespeare base-set of three plays against a counter set of 31 plays by others. The Shakespeare plays used to define Shakespeare's marker words were *The Comedy of Errors*, *The Taming of the Shrew*, and *The Two Gentlemen of Verona*. Zeta testing placed *Richard III* firmly with the plays by others. Using a larger list of 86 plays (28 of which were Shakespeare's), Zeta testing identified *Richard III* as Shakespeare's. This is not surprising: Shakespeare's marker words in this instance were determined by a substantial canon of 28 plays including all those histories and tragedies considered to be sole-authored, and the comparison dataset of "others" was 58 plays, only a two-fold difference. Genre may have been a factor too, both in Elliot and Greatley-Hirsch's *Arden* test and in Rizvi's *Richard III* test; the tragedy *Arden* was tested against a Shakespeare base set that was three-quarters comedy and *Richard III* against a purely comedic early Shakespeare. As we will see, this makes a profound difference to results.

My own testing of small base sets against larger counter sets confirms Rizvi's finding. Marlowe's full canon amounts to only a little over 100,000 words, and one or more of his plays (often *Dido*, but occasionally *Doctor Faustus*) are sometimes withdrawn from the set on the grounds they may be co-authored. It is a canon consisting almost entirely of histories and tragedies. In Craig's *1 Henry VI* tests, Marlowe's marker words were derived by comparing his canon to a large set of 130 plays of the period 1584-1641, rather than the 19 pre-1600 Shakespeare plays that Nashe's canon was set against. Craig's rationale for defining Marlowe's marker words against a large group of dramatists spanning six decades, rather than Shakespeare, was to "leave[s] open the question of whether a third writer would be a still better candidate", stating that the "ideal contrast is not between Marlowe and Shakespeare, but between Marlowe and his rival dramatists as a group" (Craig 2009, 62).

But the test conducted does not do what this rationale suggests. Firstly, dramatists such as James Shirley and Shakerley Marmion, who weren't born until several years after Marlowe's death, could hardly be classified as "his rival dramatists". The counter dataset covers an inappropriately long period (six decades), approximately ten times as long as Marlowe's productive period (1587-1593) and undoubtedly incorporating significant changes of fashionable dramatic style as well as changes in the English language. Secondly, the nature of Zeta testing does not in any way answer "the question of whether a third writer would be a still better candidate" since any potential additional co-authors would be lumped in with "others". Not having their marker words independently calculated, they would remain undetected.

Curious as to whether a small genre-skewed canon like Marlowe's could really generate reliable marker words under these conditions, I conducted an experiment in which a comparable base-set was set up for Shakespeare. Four Shakespeare tragedies whose total word-count was similar to that of Marlowe's full canon were selected for the base-set: *Coriolanus, King Lear, Antony and Cleopatra* and *Romeo and Juliet* (103,184 words). A play in the same genre, *Hamlet*, was the test text, treated as though anonymous. Segments of 2000 words were created, with rolling segmentation on the test text stepped forward by 500 words at a time. Results were analysed using the standard deviation method. In the first test, 500 words that Shakespeare favours more than others and 500 marker words others favour more than Shakespeare were generated by comparing the base-set of four Shakespeare tragedies against a counter-set of 89 supposedly well-attributed sole-authored plays by others from the period 1581-1641, not including any of Marlowe's. With this level of imbalance (a counter-base multiple of 17.2), Zeta failed to identify *Hamlet* as Shakespeare's, attributing more segments to others than to Shakespeare. The counter-set was then reduced in various ways, and the test re-run repeatedly (Figure 3).

[Figure 3]

Removing plays dated later than 1616 gave a set of 69 plays. A set of 58 plays covered a shorter period of 1590-1616 (Figure 4). Removing plays earlier than 1600 left 41 plays. Restricting the set to plays dated 1600-1610 plays created a group of 28. The test was run repeatedly, creating a new list of marker words for Shakespeare against each of these counter-sets and using *Hamlet* as the test play. But even at the reduced counter-base multiple of the 28-other set (5.57) Zeta still couldn't identify *Hamlet* as Shakespeare's.

[Figure 4]

Was *Hamlet* exceptional in having minimal stylistic affinity with the base set of four Shakespeare tragedies? To answer this question, the remaining five Shakespeare tragedies considered to be sole-authored were subjected to the same test. An overall accuracy rate was then calculated by subtracting the percentage of segments attributed to others from the percentage of segments attributed to Shakespeare. The test was 66.7% accurate at detecting Shakespeare's hand in *Julius Caesar*, perhaps because half of the base-set is Roman in its setting. "Mark" is a Shakespeare marker, being not only a verb and common noun but a major character's name, mentioned in both speech and stage directions, in *Anthony and Cleopatra*. This will lead to many segments of *Julius Caesar* (where Mark Anthony also appears) being towed towards Shakespeare. But a single word in this test only contributes 1/500th of the overall score for a segment and it is the cumulative effect that matters. Shakespeare markers with this base-set include *Julius Caesar*-friendly words like *sword, noble, war, army, enemy, soldier, general, triumph, lord* and *speak*, as well as Roman-flavoured words such as *capitol*. The test could detect Shakespeare in just under half of *Macbeth* and just over a fifth of *Cymbeline*. But its ability to spot Shakespeare in *Troilus & Cressida* and *Othello* was even more abysmal than its detection of Shakespeare's hand in *Hamlet.* The overall accuracy rate of this Zeta test across all five tragedies was a mere 12.1% (Table 5).

[Table 5].

Further reductions were made until Zeta could recognise *Hamlet* to a substantial degree. The counter-set was first reduced to cover only the plays by others from 1600-1605 (18 plays, a counter-base multiple of 3.5). The overall accuracy of Zeta's detection of Shakespeare in *Hamlet* rose to 17.5%. Finally, only the tragedies from 1600-1605 were retained: *Antonio's Revenge* by John Marston, *Mariam the Fair Queen of Jewry* by Elizabeth Cary, *Hoffman, or Revenge for a Father* by Henry Chettle, *A Woman Killed with Kindness* and *The Wonder of Women, or Sophonisba* by John Heywood and *Sejanus His Fall* by Ben Jonson. This counter-set, matched almost exactly in size to the base-set (50 segments against 49) was the only one of those tested which resulted in a modicum of accuracy in attributing *Hamlet* (Figure 5).

[Figure 5]

This experiment demonstrates that with early modern texts, when base sets are around 100,000 words, a large disparity between base and counter set sizes (3.5 and higher) leads to wildly inaccurate results. Since none of Craig's 2009 Zeta tests on Marlowe, Nashe or Kyd had a counter-base multiple below 3.7, it is highly unlikely that any of them are accurate.

Effect of matched small datasets on Zeta results

However, with small datasets like these (100,000 words), matching counter-set size to the base-set doesn't necessarily improve Zeta's reliability. In a further experiment designed to test whether matching dataset sizes would maintain or improve validation results, Burrows and Craig's Shakespeare dataset (at 148,000 words, approximately 1.5 times

Marlowe's canon) was reduced to the four history plays *Richard II, Richard III, Henry V* and *King John* (113,000 words). This was compared with Marlowe's full canon to derive Shakespeare and Marlowe marker words, and the two sets of seven Marlowe and seven Shakespeare plays were re-tested under the same conditions as the previous validation test with one exception. Because of my concern that withdrawing *Richard III* or *King John* for their Zeta analysis from Shakespeare's four would destroy the 1:1 dataset balance I was attempting to test (each of them being a larger proportion of the whole set than any of Marlowe's plays), when one of these plays was tested, *2 Henry IV* was substituted into the base-set in its place.

With these matched size (and genre) datasets, Zeta's accuracy fell to 40.2% for Shakespeare and 49% for Marlowe. Making both base and counter-sets approximately 100,000 words led to erratic Zeta results with success varying from over 60% (*1 Henry IV* and *King John* in this Marlowe-comparison test, *Hamlet* in the previous tragedies-by-others test) to zero (*The Taming of the Shrew*) (Table 6). The first time these tests were run, I made the mistake of substituting the tragedy *Julius Caesar* in the base set when *King John* and *Richard III* were tested rather than the history play *2 Henry IV*. When Julius Caesar was the fourth play, *King John* came in the bottom three for accuracy, testing as 28.7% Shakespeare's. With *2 Henry IV* as the substituted play, *King John* came top in the table, testing as 77.7% Shakespeare's. This demonstrates the fragility of Zeta test results with small datasets. It is also one of several indicators of the strong influence of genre on Zeta.

[Table 6]

Effect of genre on Zeta results

[Figure 6]

Zeta's inability to detect Shakespeare's hand in *The Taming of the Shrew* under the conditions of this particular sub-method is a stark illustration of a long-suspected weakness of Zeta (Figure 6). Many researchers suspect that genre unduly influences the results of lexical word tests, and this belief is the reason why function words tests, supposedly less influenced by genre, are often employed.

In response to Joseph Rudman's claim that many studies show that "genre trumps authorship" Hugh Craig ran some tests on the plays of Shakespeare and four of his contemporaries that compared the frequency of 100 commonly-used words (largely function words) across dramatic sub-genres (comedy, tragedy, history) and between authors (Rudman 2016, 318; Craig 2017).[10] He concluded that "[t]he experiment set up to estimate genre difference compared to author difference shows that author difference is much greater -- *at least with these variables, these genres, these authors and these play sets*" (156, my italics). I would add to that caveat 'these particular tests'. The test Craig used was a statistical tool called a t-test.[11] T-tests have been used successfully to discover where one author took over from another in a co-authored 19th century novel where both authors are known (Hoover 2010). However, they are not used to determine the attribution of Early Modern plays, the context of Rudman's statement, and are not the tests Craig uses in his own attribution work. He recognises that "[w]ith other kinds of evidence, the picture might be different. No doubt more lexical words would reflect genre more strongly, for instance" (163). Since the main test Craig uses for detailed attribution work is Zeta, a lexical word

[10] Rudman stated "It has been shown in many studies that genre trumps authorship – there is a greater stylistic difference between one author in different genres than between two authors writing in the same genre." As Craig notes, he did not cite the studies.

[11] Craig selected 100 commonly used words and a "hetero-scedastic version [of the t-test] known as Welch's t-test". This results in a probability (0-1) that "two sets [of data] came from the same parent population." He then counted how many of the 100 words had statistically significant differences in frequency of use ($p<0.01$).

test, it is surely *this* test that should be used to investigate whether genre (at least in some cases) trumps authorship.

The hard zero that arises from Zeta-testing *The Taming of the Shrew* against two datasets in a different genre (one of them by the presumed author) is a good illustration of Zeta detecting lexical markers that have nothing to do with authorship. Whereas none of *The Shrew*'s segments are attributed to Shakespeare when the marker words are derived from comparing 100,000 words of his histories with Marlowe's canon, Zeta has no problem attributing *The Shrew* to Shakespeare when the marker words are derived from his comedies. When an identical test was run with a base set comprising *Much Ado About Nothing, Love's Labour's Lost, The Comedy of Errors, The Merry Wives of Windsor* and *All's Well That Ends Well* (102,171 words), attribution of *The Shrew*'s segments to Shakespeare leapt from zero to 91.6% (Figure 7).

[Figure 7]

The implications of this for authors with small genre-skewed canons is profound. We know that Marlowe's two *Tamburlaine* plays had comic scenes that the printer, Richard Jones, stripped out, and that there are comic scenes and humorous elements in *Doctor Faustus* and *The Jew of Malta,* but his surviving attributed canon is of a markedly tragical-historical bent. Consequently, if Marlowe had a hand in one of the comedies of this period, Zeta would never be able to reveal it.

Revisiting the Henry VI attributions

3 Henry VI

Burrows and Craig established that a particular set of eight early Shakespeare histories and tragedies (against Marlowe's full canon) appears to lead to good levels of attribution accuracy (particularly in Shakespeare's case), and my own tests confirm this

(Table 3). With the addition of the new standard-deviation interpretation method that allows some segments not to be attributed either way, *3 Henry VI* was Zeta-tested using these datasets. Burrows and Craig, using the dubious bisector line method, attributed 108 segments of the 1623 Folio version of *3 Henry VI* almost equally between the two authors, 54.4% to Shakespeare and 45.6% to Marlowe. The standard-deviation method reveals that the vast majority of *3 Henry VI*'s segments fall in no man's land, too far from either author's centroid to be matched to them.[12] 18.9% can be attributed to Shakespeare, 9% to Marlowe, and in the remainder (72.1%), the method is unable to distinguish between the two (Figure 8). With balanced datasets, where Zeta struggles to tell Marlowe from Shakespeare to an even more marked degree, Marlowe's portion of 9% is retained, while Shakespeare's drops to 1.8%.

[Figure 8]

These results could not be more inconclusive. Zeta suggests Marlowe had a small hand in *3 Henry VI*, but with so little attributed to Shakespeare, it's hard to know what we can conclude from the results. And we can now understand that the tests used to rule out other playwrights were flawed. Burrows and Craig ruled out Kyd, Greene and Peele by a variety of Zeta tests: 1000 lexical markers, then 500 function word skip bigrams, then an undeclared number of lexical word skip trigrams (Burrows and Craig 2017, 204-10).[13] All data was analysed using the inherently faulty bisector line method. Across these tests, *3 Henry VI* segments clustered with "others" in the case of Greene, Kyd and Peele but merged with the Marlowe cluster. Indeed, in both function and lexical word Zeta tests,

[12] So far as 'style' can really be defined by favouring certain individual words over others. Though I have strong reservations about such a definition I will continue to use 'style' in this manner for convenience.

[13] A bigram is two adjacent words e.g. the first bigram in "the rain in Spain" is "the-rain". A function word skip bigram is made of adjacent function words, skipping lexical words between them e.g. "the-in". Trigrams are created from three words.

even some segments of *3 Henry VI* previously designated as “Shakespearean” clustered with Marlowe segments (Craig and Burrows 2012, 61; Burrows and Craig 2017, 209 Fig 11.9). Given the size of Greene, Kyd and Peele’s canons, and what we now know about the effect of counter-base size disparity when Zeta-testing canons of around 100,000 words, and the additional inaccuracy of Zeta tests using two small datasets (e.g. deriving marker words for Kyd by comparing his canon to Marlowe’s), all these results should be discarded. It is questionable whether Greene, Kyd and Peele have dramatic canons large enough for a reliable Zeta sub-method to be devised and validated. Only the Zeta sub-method that was independently validated to be able to tell Shakespeare’s plays from Marlowe’s is worthy of duplication, and even that shows concerning fluctuation in accuracy between plays.

1 Henry VI

What about the other two *Henry VI* plays? In Craig’s 2009 tests on the first two plays, Marlowe’s marker words were derived against a dataset 23.4 times the size of his corpus, and as has been demonstrated, *Hamlet* (and indeed all other Shakespeare plays I have tested) cannot be recognised as Shakespeare’s under these conditions, so we can assume the results for Marlowe were equally invalid.

The fact that Craig’s *1 Henry VI* tests do not consider issues of genre is a further cause for concern. Craig gives a list of words that appear in the middle Joan sequence of *1 Henry VI*, and are in the top 50 of Marlowe marker words: *gold, arms, realm, pride, slain, sword, golden, overthrow, death, base, damned, foe, field, yield, cruel, stay, conquering, hell, countries, words, terror.* A simple word count for each of these words across Marlowe’s canon is instructive (Table 6)

[Table 6]

The two Zeta tests on *The Taming of the Shrew* have demonstrated that genre is critical and will skew results on lexical word tests, especially where the author's corpus is small, as it is here. Many of the Marlowe marker words present in the middle "Joan" section are associated with warfare and tyranny — *arms, realm, sword, overthrow, death, foe, field, yield, cruel, conquering, terror*. The majority of the marker words are strongest in the three plays that are full of battles – the two parts of *Tamburlaine*, chiefly, with *Edward II* in a significant third place. This is hardly surprising, given that Joan is at war with the English. Since the three Joan sequences selected for this test are chiefly concerned with warfare, they will tally with the words in Marlowe's canon also concerned with warfare. The marker word "hell" (and to some extent "damned') are present in the two *Tamburlaine*s (with their devilish protagonist) but chiefly derive from *Doctor Faustus*, which concerns a man damned to hell. One wonders what words a writer should use to address their particular subject matter other than death, arms, sword, hell, etc.

"Countries", we are told, is in the top 50 of 500 Marlowe marker words, but Marlowe uses it only twice across his entire corpus of six plays. Zeta is meant to measure words of middling frequency; it supposedly "addresses word-types that occur with some consistency in one sample batch but not in another comparable batch" (Craig and Burrows 2012, 57). Two uses across the whole corpus surely doesn't fit the description of "middling frequency". This is another illustration of how unreliable Zeta can be when testing authors with a small corpus. But it turns out that the word shouldn't have appeared in the list of *1 Henry VI* words that are also in Marlowe at all. *1 Henry VI* doesn't contain the plural "countries"; Joan la Pucelle is using the possessive "country's".

"Death" is a word which is strong across the entire Marlowe canon. But the accepted Marlowe canon minus *Dido* - which was left out as possibly co-authored - is entirely comprised of histories and tragedies, or in *The Jew of Malta*'s case perhaps a tragic farce, so this is unsurprising. "Death" was picked up as a marker word for Marlowe against

130 single author non-Marlowe plays, including Jonson's city comedies, Marston's satires, and all kinds of frippery.

If we compare Marlowe's historical and tragic plays with Shakespeare's historical and tragic plays "death" does not remain a significant Marlowe marker word. I conducted a word count of the "Marlowe marker works" supplied by Craig as they appear in 30 of Shakespeare's 36 First Folio plays (leaving out the *Henry VI* plays, H*enry VIII, Timon of Athens* and *Titus Andronicus* as probable co-authored plays). It becomes clear that many of the words supposedly significant for Marlowe are also significant for Shakespeare. There is no play in Shakespeare's corpus where the word "death" doesn't occur, and he used it a total of 664 times in the 30 plays. *Romeo and Juliet* and *Richard III* share the highest usage count at 69 apiece. If you remove the plays the Folio classifies as comedies to genre-balance the corpora, their average usage is near-identical: 30 uses per play for Shakespeare, 29 for Marlowe. "Sword" is a similar case.

The "Marlowe marker words" favoured by Shakespeare are not confined to those that are genre-specific. Shakespeare is every bit as fond of "stay" and "words" as Marlowe seems to be, using "stay" 316 times in the 30 Folio plays and "words" 298 times. Only a handful of the words listed — "realm", "overthrow", "foe", "yield" and "terror" — fall outside the ranges found for these words in Shakespeare's Folio plays.

Just a brief look at a few of these "top 50" Marlowe marker words has illustrated why we might have concerns about the method when it is not conducted under conditions that take genre into account. One may argue that with a list of 500 marker words, problems with individual words will be ironed out, but without running our own tests, we cannot know how seriously the results have been affected by the choices made.

The tests Craig ran to investigate Marlowe's co-authorship of *1 Henry VI* were not only designed without consideration for genre differences and the effect of counter-base dataset size imbalance; they were incomplete. Marlowe's style (such as a collection of words and the inappropriate counter-set defines it) was only sought in the three sequences relating to Joan La Pucelle rather than the whole play (Craig 2009, 61-2).[14] As Rizvi has pointed out, the fact that the two groups (Marlowe and non-Marlowe) cluster in different parts of the resultant graph is "not a research finding but merely what we expect from the definition of the method" (Rizvi 2019, 403). Yet there is some significant cross-over between the two clusters of Marlowe and non-Marlowe play segments, unlike the separation seen for Nashe's prose when compared with drama by others (Craig 2009, 63 fig 3.9). As Craig and Kinney describe in their "Methods" chapter, this overlap means that "the diversity of vocabulary patterns within each of the two groups, and the degree of common ground between the two, have defeated our attempts to make a clear separation" (Craig and Kinney Shakespeare, Computers 2009, 19). The three Joan sequences of *1 Henry VI* fall almost exactly on the bisector line: the early sequence slightly above it (closer to the centre of the "non-Marlowe" cluster) and the mid and late sequences slightly below it (closer to the centre of the "Marlowe" cluster), but the difference is minimal. Without the bisector line drawn onto it, the authorship of the segments is by no means clear-cut.

Using the datasets that we know to have a high level of reliability for distinguishing the two authors for most (though not all) of their sole-attributed plays, the whole of *1 Henry VI* was tested and analysed using the standard-deviation method (Figure 9). By this method, Marlowe claims a single segment early in Act 1, corresponding with I.ii.22-150, which was included in Craig's early Joan sequences. Most of Act 2 (which includes the Temple Garden scene) is attributed to Shakespeare. Marlowe gets the second half of Act 3, and early Act 4. Craig's "middle Joan sequences" are contained within the second half of Act 3, but Marlowe's favoured words are detected continuing beyond them. Marlowe is also

[14] The rationale for this decision is that these are sections that Thomas Merriam thought might have been written by him on the basis of certain verbal parallels and PCA function word tests. See Merriam, T. (2002). "Faustian Joan." Notes and Queries **49**(2): 218-20.

attributed almost all of Act 5, far more than Craig's "late Joan sequences" which are included within. Talbot's last battle scene (4.vii), which Craig attributed to Shakespeare, cannot be confidently attributed to either author. This is not to say that it was not written by one of them. As with the rest of the unattributed segments of the play, it only means that this particular Zeta sub-method, which, unlike Craig's 2009 test, is validated for the two authorial canons and analysed by a mathematically sound method, cannot tell the two authors apart. Overall, 14.3% of *1 Henry VI* was attributed to Shakespeare and 22.4% to Marlowe, with the remaining 63.3% of the play of indeterminate authorship. This may be due to other authors, such as Kyd, having a hand in the play, but is also likely to rest quite heavily on Zeta's limited ability to tell Marlowe from Shakespeare in certain plays.

[Figure 9]

2 Henry VI

A similar story applies to *2 Henry VI*. However, Craig's attribution of (most of) Act 4 of *2 Henry VI* to Marlowe is not achieved by Zeta alone, but by combining Zeta results with the results of function word PCAs (Principal Component Analyses), and only attributing to Marlowe those segments where the two tests concur. There is much that might be said about the conflicting results of the two methods, and indeed the need to bring in PCAs at all for this play, but our focus here is what our validated, genre-balanced Zeta sub-method might tell us.

[Table 8]

Analysing *2 Henry VI* using Burrows and Craig's 2017 Zeta validation datasets and the standard deviation method (at three standard deviations), 23.9% of the segments are given to Shakespeare, and 17.1% to Marlowe (Table 8). Marlowe's "marker words" are detected in five segments towards the end of Act 1 and fourteen segments in the second half of Act 4, corresponding to a large part of the Jack Cade scenes. This doesn't mean that Marlowe didn't write the whole of Act 4, only that this more stringent Zeta analysis method can't distinguish Marlowe from Shakespeare so easily. Those parts of *2 Henry VI* given to Shakespeare include two segments at the start of Act 2, a run of twenty-two segments covering the end of Act 2 and the first half of Act 3, two segments near the end of that act, and the final segment. As with the other two plays in the trilogy, the majority of segments fall, stylistically, into the no man's land between Marlowe and Shakespeare (Figure 10).

[Figure 10]

The Jew of Malta and other Zeta anomalies

As discussed, two of Marlowe's plays – *Doctor Faustus* and *The Jew of Malta* – behave similarly under these conditions. *The Jew of Malta* goes further, with a significant proportion of it being allocated to Shakespeare. Burrows and Craig explained the "spectacular failure" of Zeta by saying "the test can be deceived by a play like *The Jew of Malta*, which evidently departs from the Marlowe style as represented by the other six plays, and which might be regarded as more confessional, more farcical, and more comic than they are, though still Marlovian" (212). This is to admit the genre-dependent nature of lexical word tests and the genre-skewed nature of Marlowe's canon. But it raises the question of why certain of Shakespeare's comedies are identified as 100 per cent Shakespeare, even by the standard deviation method, when the baseline Shakespeare subset specifically excluded comedy (Table 3).

What does the anomalous *Jew of Malta* result reveal about this particular Zeta test, about the nature of the Shakespeare canon, or about the relationship between Marlowe's word choices and Shakespeare's? By the standard deviation method, Marlowe is allocated just over a third of the play, and Shakespeare a quarter, with the rest of the play of undetermined authorship (Figure 11). This is not to say that the entire play isn't Marlowe's, only that this particular Zeta sub-method cannot detect that the play is fully Marlowe's. The same sub-method's categorisation of *Henry V* as only two third's Shakespeare's should lead us to question the test before we question the authorship of the play. Nevertheless, there is something worth investigating here.

[Figure 11]

The earliest text we have for *The Jew of Malta* dates from 1633 and includes a prologue by Thomas Heywood, raising the possibility of post-1592 revisions (by Heywood or others) in the rest of the text. The work of Thomas Merriam and Robert Matthews, mapping individual Marlowe and Shakespeare plays against stylistic markers, replicated the anomaly found two decades later by Burrows and Craig (Merriam and Matthews 1994, 4). In their results graph, *The Jew of Malta* shows as separate from Marlowe's canon, sitting at the far-Shakespeare edge of a cluster of Shakespeare plays; out-Shakespearing Shakespeare in almost every regard.[15]

Though there are likely to be as many if not more issues with Merriam and Matthew's methods as with the endlessly variable Zeta, it's possible that both sets of results are nevertheless revealing useful information. What they do *not* show is a distinct separation between the two authors that can be consistently and reliably revealed. And though *The Jew of Malta*'s anomalous qualities might be explained by a very late revision, the revision would need (according to Merriam and Matthews' tests) to have been by someone who wrote with the same stylistic markers as Shakespeare in his maturity.

[15] It has the lowest (and a negative) "Marlowe signal'; only *Cymbeline* and *Timon of Athens* have a stronger "Shakespeare signal".

Zeta results suggest more of a continuum in word choices, though this may be a visual illusion created by the design of the test (and the resulting graph). *The Jew of Malta* behaves quite similarly to the *Henry VI* plays, in that it appears to present a wide swathe of dots connecting the two authors' data clouds. Also, we should recall that we are attributing according to a very generous three standard deviations from author centroids. At two standard deviations, *The Jew of Malta* is so stylistically deviant (according to Zeta) that it cannot be attributed to either author; the same is true for *3 Henry VI* and very nearly true for the other two plays in the trilogy.

Whether these and other Zeta results tell us more about the text or the tests, is open to question. Certainly, these latest results suggest that in the case *The Jew of Malta, Doctor Faustus*, the three *Henry VI* plays, and to a lesser extent, *Henry V, Antony and Cleopatra, The Two Gentlemen of Verona* and *The Tempest*, even well-validated Zeta sub-methods struggle to distinguish Marlowe's word choices from Shakespeare's.

Summary

The more rigorous method for analysing Zeta results introduced in this article creates results that have a higher degree of uncertainty than the bisector line method, but that is no bad thing. That uncertainty reflects the data itself, which in many Zeta results, is gathered in the no man's land between two authorial clusters, either in a discrete cloud or in something resembling a stylistic continuum. In determining whether to attribute segments of a test play to the base author if it is within two or three standard deviations, we have discovered that only the most generous of these measures, covering 99% of the standard distribution curve, leads to acceptable success rates, suggesting that Zeta is far more fragile than Gabriel Egan's "most powerful … authorship tool currently available" would imply

(Egan 2017, 45). Nevertheless, given the other methods at our disposal, he may still be correct.

The experiments conducted here have demonstrated the importance of proper validation for each Zeta sub-method, to ensure it will attribute to each author being tested, with reasonable accuracy, works accepted as theirs. However, even if the overall percentage success rate is impressive, the presence of significant anomalies for a particular sub-method must be noted, with the understanding that the test play may be as difficult to categorise as these anomalous plays are.

It is now clear that base-sets as small as 100,000 words are unlikely to give reliable Zeta results with counter sets at a multiple of 3.5 or more (no matter which authors are being compared), and that 1:1 matched size datasets at this base-set size are also unreliable. All Zeta tests results with small base-sets, with either 1:1 size-matched counter sets or high levels of counter-base size disparity should, therefore, be discarded. This includes all of Craig's tests on the first two *Henry VI* plays with regards to Kyd, Nashe, Greene, Peele and Marlowe.

Genre influence on Zeta appears to be complex and needs further investigation. Comparing matched-genre base and counter-sets seems intuitively correct given that Zeta will otherwise be detecting genre markers rather than authorship markers, and this was presumably Burrows and Craig's intention in devising the validation subset of eight Shakespeare tragedies and histories to match Marlowe's canon in their 2017 validation tests. However, under these conditions the test is considerably more reliable for some plays (of both authors) than others, with confidence-denting results. The genre similarity of both datasets can 'tow' significant proportions of certain plays into a middle ground that makes these segments appear to belong to neither author, when this may not be (or is presumably not) the case.

Despite Zeta being apparently able to distinguish Shakespeare texts from Marlowe's with an accuracy of 91.6% (Table 3), tests on the *Henry VI* plays shed more light on the test itself than on the composition of the plays. Between 59% and 72.1% of each of these plays sits so comfortably *between* Shakespeare's style and Marlowe's (at least as measured by favoured word frequency) that Zeta is unable to decisively attribute the major part of any of these plays to one author or the other. Though other authors, with canons too small to test by this method, may have contributed in some way, it is unlikely that nearly three-quarters of *3 Henry VI* was written by authors who are neither Marlowe nor Shakespeare. The result is more likely to mean that in 72.1% of the play, the difference between what the test has determined is either author's style (as far as style can be measured by word choice) is too close for Zeta to call. Zeta's difficulty in differentiating between the two writers is apparent in later plays too, notably in *Henry V, Antony and Cleopatra*, and *The Tempest*. These results do not require that we consider these plays as only partially written by Shakespeare.

Is authorial influence a factor here? Marlowe's influence is recognised to be strong across the entirety of Shakespeare's output, and has been the subject of at least one full-length study (Logan 2007). The Zeta method is formulated to seek differences in word-use so where there is a great deal of similarity (through influence or otherwise) these words will not end up as markers for either author. As a result, words that are not very prevalent (as we saw with 'countries') can end up in the top 50 marker words for an author, leading to fragile results. The effect of influence on Zeta would be useful to study, but in Early Modern drama it is hard to find datasets large enough to test. You might, for example, want to know whether there are linguistic similarities between Marlowe's *Tamburlaine* plays and Robert Greene's *Alphonsus, King of Aragon*. But at four plays, Greene's dramatic corpus is even smaller than Marlowe's and therefore, as we have seen from this study, not large enough for Zeta to produce meaningful results.

These indistinct results for the *Henry VI* plays also leave open the possibility that any one of the plays might be written wholly by the same author. This is because the separation on Zeta graphs 'is not a research finding but merely what we expect from the definition of the method' (Rizvi 2019, 403). Where Zeta marker words are determined by comparing two sets of texts against each other (rather than comparing each, separately, against a third dataset) it will necessarily force a separation that may have nothing to do with author identity. Where nine early Shakespeare plays are set against nine late Shakespeare plays, a Zeta test on *Henry V* results in 81% of the segments falling into no man's land, yet the person who wrote the play was (as far as we know) the author of the texts in both groups.[16] With enhanced understanding of how Zeta functions under various test conditions, we can only conclude that any one of the *Henry VI* plays may or may not be co-authored.

Computational stylistics really does need "big data" and few of Shakespeare's suspected co-authors can provide it. Though Zeta has more limitations and fragility, when applied to small canons, than previously appreciated, it might yet be a useful tool for interrogating questions of authorship when responsibly applied, with genre compensation, dataset size control, and each sub-method fully validated for the authors involved. A great deal more testing will be needed to discover the extent of its capacities and limitations when used on early modern plays. A validated and mathematically sound Zeta analysis of the *Henry VI* plays is apparently able to detect words Marlowe favours more than Shakespeare in all three plays. This might be taken as confirmation of a long tradition of attributing these plays, at least in part, to Marlowe. But since neither author can confidently

[16] 1000 marker words generated. Early play base set: *Romeo & Juliet, Two Gentlemen of Verona, The Taming of the Shrew, Richard III, The Comedy of Errors, Love's Labour's Lost, Richard II, A Midsummer Night's Dream, King John. Late play counter set: Troilus & Cressida, Hamlet, Othello, King Lear, Antony & Cleopatra, Coriolanus, The Winter's Tale, Cymbeline, The Tempest.* 2000-word segments with test text (*Henry V*) subjected to rolling segmentation in 200-word increments.

be attributed more than a quarter of the segments, we may be learning more about the tests than the texts.

Table 1. A replication of Burrows and Craig's 2017 Zeta validation test, where marker words for Marlowe and Shakespeare are derived from comparing the seven plays of Marlowe's canon against a set of six Shakespeare histories and two Shakespeare tragedies dated before 1600. Each play is withdrawn from the underlying set before being tested. Assignations to Shakespeare or Marlowe are based on a segment being within either two or three standard deviations from authorial base set while being outside three standard deviations of the counter set. Results for Marlowe.

	Number	Identified as Marlowe		Identified as Shakespeare		Attributed to Author	
	of Segments	2-STDEV	3-STDEV	2-STDEV	3-STDEV	2-STDEV	3-STDEV
Doctor Faustus (1604)	49	0	12	0	0	0.0%	24.5%
Edward II	96	13	77	0	0	13.5%	80.2%
Jew of Malta	82	0	29	0	21	0.0%	35.4%
1 Tamburlaine	79	61	79	0	0	77.2%	100.0%
Dido Queen of Carthage	59	11	53	0	0	18.6%	89.8%
2 Tamburlaine	81	73	81	0	0	90.1%	100.0%
Massacre at Paris	42	9	23	0	0	21.4%	54.8%
TOTAL	**488**	**167**	**354**	**0**	**21**	**34.2%**	**72.5%**

Table 2. A replication of Burrows and Craig's 2017 Zeta validation test under the same conditions as Table 1, using two variations of the new attribution method. Results for Shakespeare.

	Number	Identified as Shakespeare		Identified as Marlowe		Attributed to Author	
	of Segments	2-STDEV	3-STDEV	2-STDEV	3-STDEV	2-STDEV	3-STDEV
Love's Labour's Lost	96	63	84	0	0	65.6%	87.5%
Merchant of Venice	97	75	97	0	0	77.3%	100.0%
King John	94	63	91	0	0	67.0%	96.8%
Comedy of Errors	64	24	54	0	0	37.5%	84.4%
Richard III	136	26	104	0	0	19.1%	76.5%
Taming of the Shrew	95	25	90	0	0	26.3%	94.7%
1 Henry IV	114	98	114	0	0	86.0%	100.0%
TOTAL	**696**	**374**	**634**	**0**	**0**	**53.7%**	**91.1%**

Figure 1. Zeta scatter plot for *Doctor Faustus*, using the Burrows and Craig 2017 validation datasets (212), comparing the new standard deviation method with the bisector line method. For comparability across tests, the base set is always Shakespeare and the counter set always Marlowe.

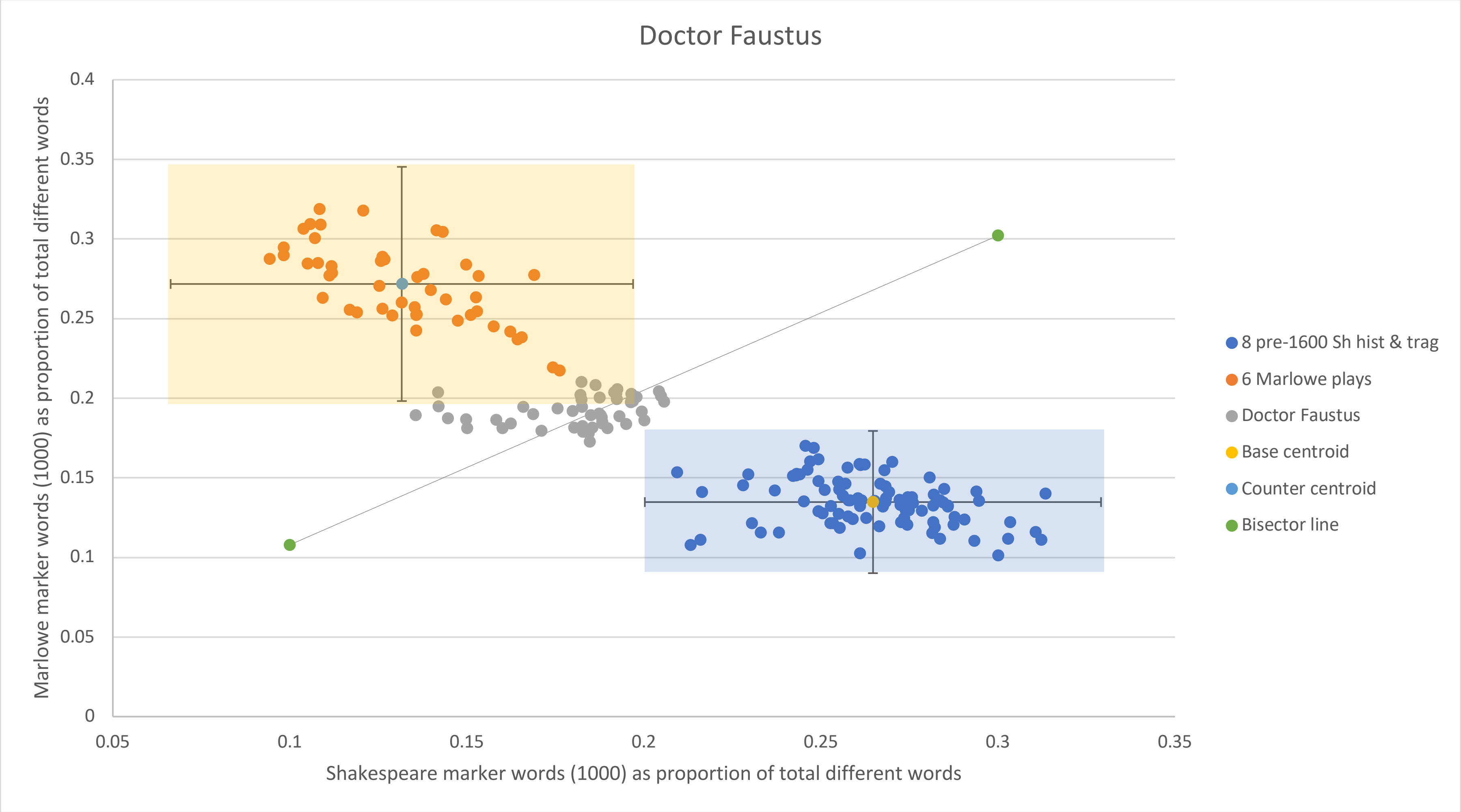


Table 3. 30 Shakespeare plays (considered by most scholars to be largely sole-authored) tested using Zeta with the same datasets, for Marlowe and Shakespeare, that Burrows and Craig used in their validation tests, using the new attribution method. Segments are assigned to Shakespeare if they are within three standard deviations of the Shakespeare base set and outside three standard deviations of the Marlowe counter set.

Play name	Genre	Total Segments	Attrib. to Shakespeare	Accuracy %
The Tempest	C	74	54	73.0%
The Two Gentlemen of Verona	C	76	56	73.7%
Comedy of Errors	C	64	54	84.4%
A Midsummer Night's Dream	C	73	62	84.9%
Love's Labour's Lost	C	96	84	87.5%
All's Well That Ends Well	C	106	100	94.3%
Taming of the Shrew	C	95	90	94.7%
Merchant of Venice	C	97	97	100.0%
Much Ado About Nothing	C	96	96	100.0%
As You Like It	C	99	99	100.0%
Measure for Measure	C	98	98	100.0%
Twelfth Night	C	90	90	100.0%
The Merry Wives of Windsor	C	98	98	100.0%
The Winter's Tale	C	115	115	100.0%
Henry V	H	121	79	65.3%
Richard III	H	136	104	76.5%
Richard II	H	101	90	89.1%
King John	H	94	91	96.8%
1 Henry IV	H	114	114	100.0%
2 Henry IV	H	122	122	100.0%
Antony and Cleopatra	T	114	77	67.5%
Cymbeline	T	128	115	89.8%
Macbeth	T	76	69	90.8%
Coriolanus	T	129	118	91.5%
Julius Caesar	T	88	83	94.3%
Troilus and Cressida	T	120	114	95.0%
Romeo and Juliet	T	112	109	97.3%
King Lear	T	120	117	97.5%
Hamlet	T	142	139	97.9%
Othello	T	122	121	99.2%
		3116	**2855**	**91.6%**

Figure 2. Zeta scatter plot for Shakespeare's *Henry V*, using Burrows and Craig's 2017 Marlowe and Shakespeare Zeta validation datasets (Burrows and Craig 212).

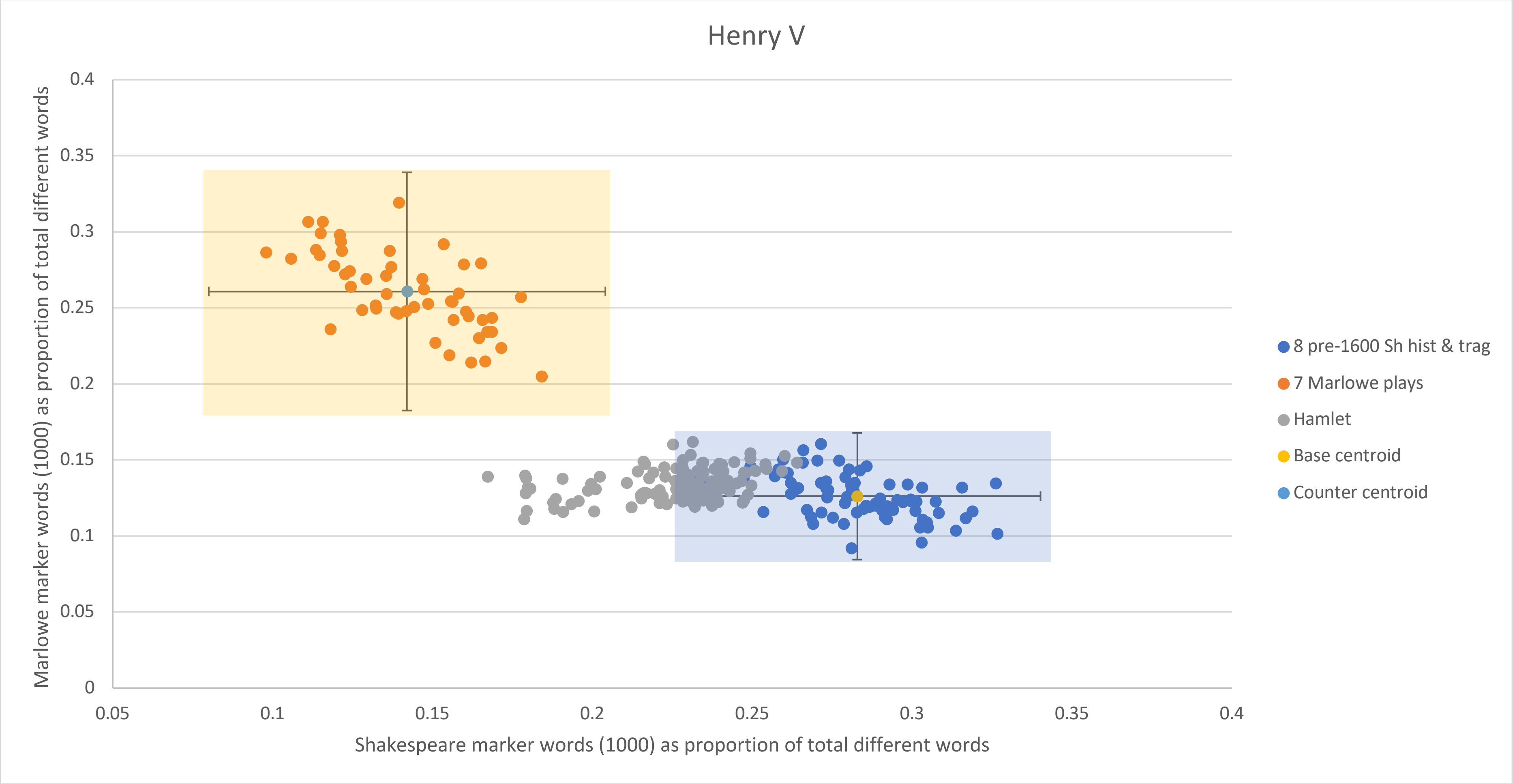


Table 4. Zeta method variations used to analyse possible co-authors of *1 Henry VI* in Craig (2009). The text says the segment size was 4000 words for the first Nashe test (54), but the axis label for the results graph of this test says 2000 words (55). I have assumed the text is correct.

Author	Segment Size (words)	How much of 1H6 tested	1H6 divisions	Base set	Base set size (est. words)	Counter set	Counter set size (est. words)	Counter/base multiple
Nashe	4000	All	Act	3 prose works	100,000	49 pre-1600 plays by others	852,000	8.5
Nashe	2000	All	Act	3 prose works	100,000	17 Pre-1600 Shakespeare plays	369,946	3.7
Kyd	2000	All	2000-word segments	2 plays	34,000	134 plays by others	2,562,000	75.4
Kyd	2000	All	2000-word segments	2 plays	34,000	48 pre-1600 plays by others	832,000	24.5
Kyd	2000	All	2000-word segments	3 plays	50,000	134 plays by others	2,562,000	51.2
Kyd	2000	All	2000-word segments	3 plays	50,000	48 pre-1600 plays by others	832,000	16.6
Marlowe	2000	3 'Joan' sections	1800-2350-word sections	6 plays	100,000	130 plays by others	2,340,000	23.4

Figure 3: Effect of counter/base set size disparity: testing *Hamlet* against base of four Shakespeare tragedies with various counter-set sizes.

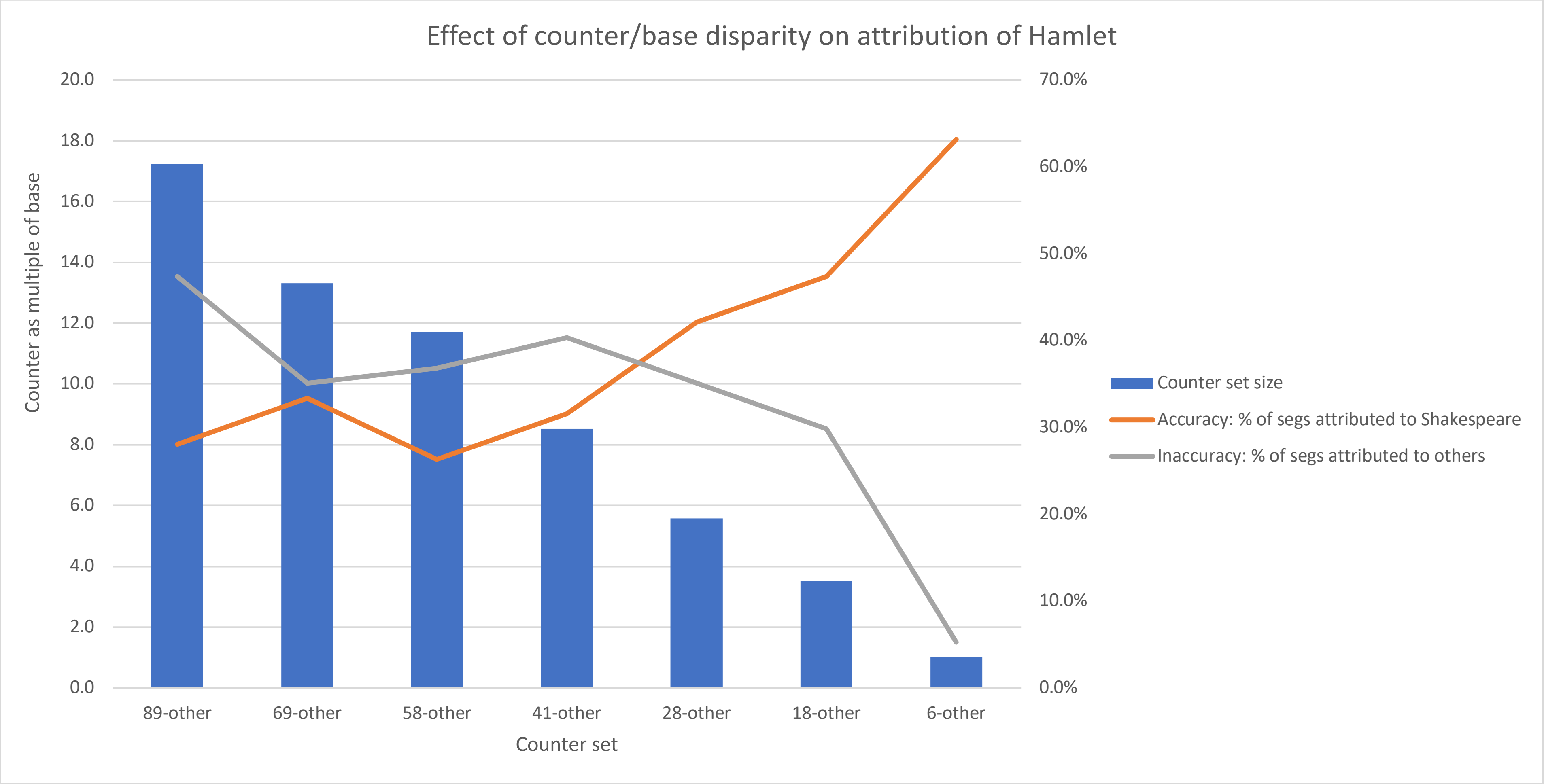


Figure 4. *Hamlet* poorly-recognised: a Zeta graph of *Hamlet* compared with a small base-set of Shakespeare tragedies and a large counter-set.

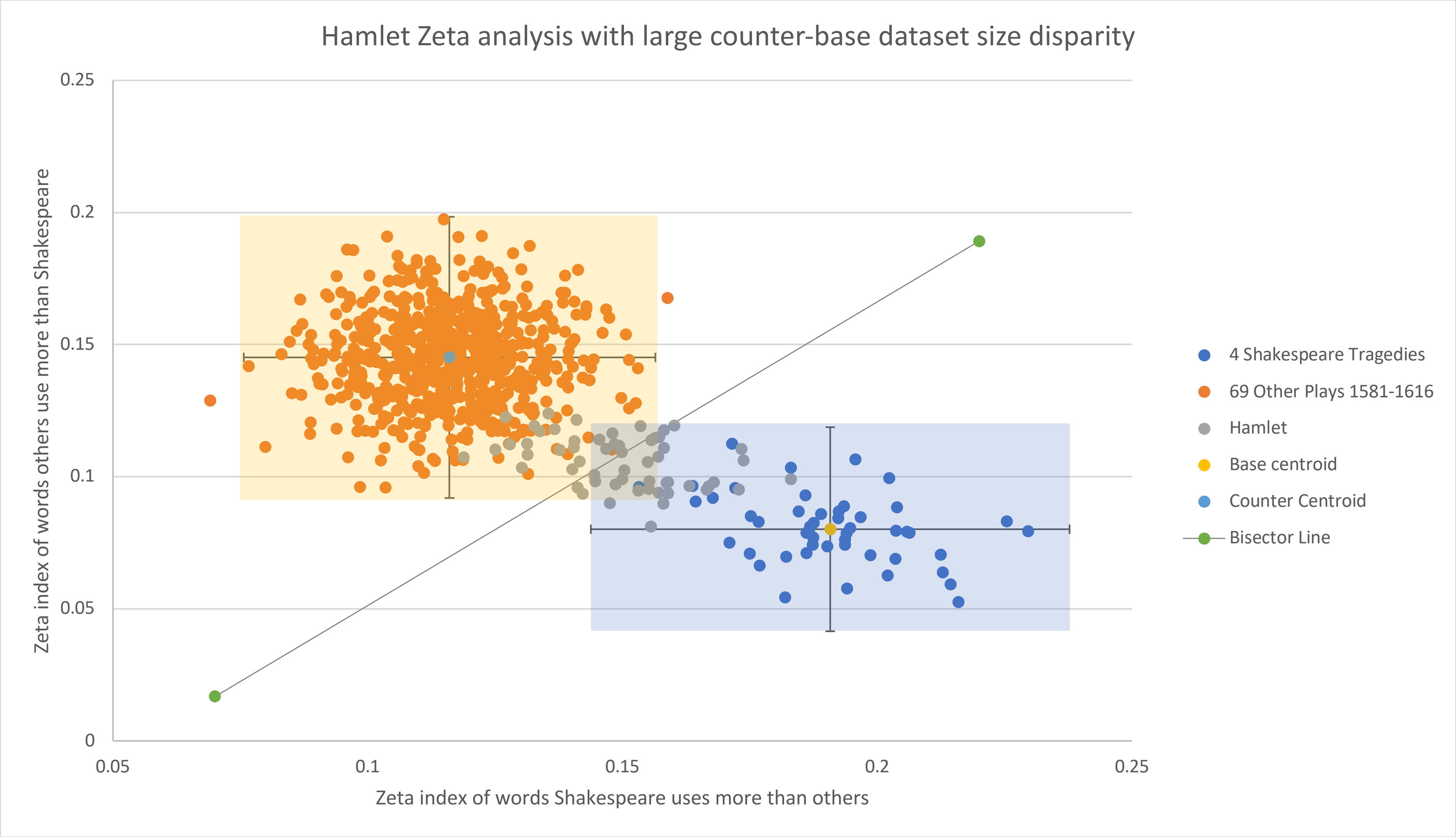


Figure 5. *Hamlet* well-recognised: a Zeta graph of *Hamlet* compared with a similar size base and counter set.

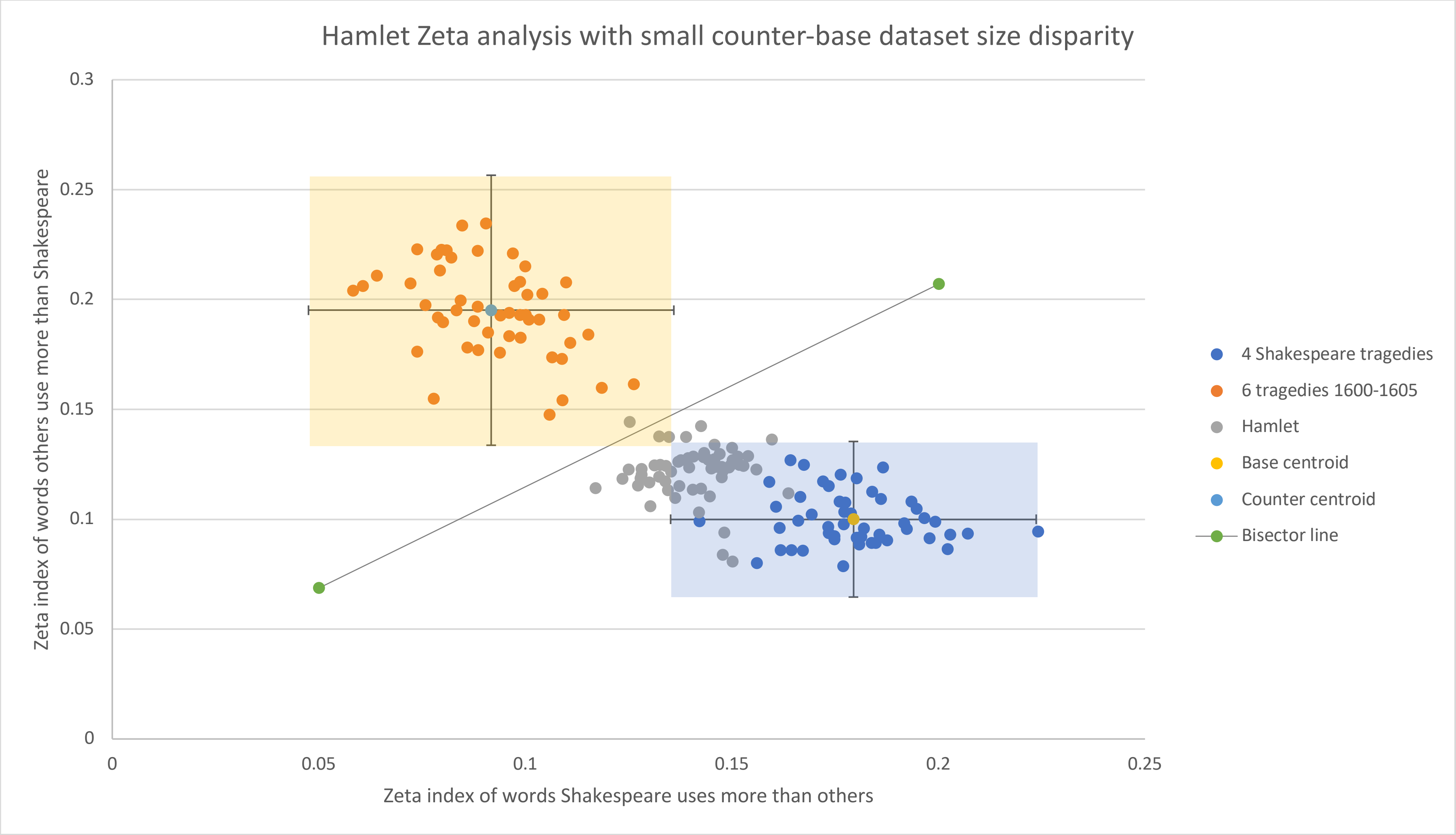


Table 5: Effect of counter/base size disparity: accuracy of a Zeta test detecting Shakespeare marker words in six Shakespeare tragedies with the remaining four tragedies as the base-set and 28 plays by others from 1600-1610 as the counter-set.

	Segments Attrib to Sh	Segments Attrib others	Total Segments	Accuracy	Inaccuracy	Accuracy - Inaccuracy
Hamlet	24	20	57	42.1%	35.1%	7.0%
Julius Caesar	25	1	36	69.4%	2.8%	66.7%
Troilus & Cressida	8	26	48	16.7%	54.2%	-37.5%
Cymbeline	22	11	52	42.3%	21.2%	21.2%
Macbeth	18	3	31	58.1%	9.7%	48.4%
Othello	10	13	49	20.4%	26.5%	-6.1%
	107	74	273	39.2%	27.1%	12.1%

Table 6. Effect of matched (100,000-word) datasets on Zeta's ability to tell Marlowe and Shakespeare apart, using the plays originally tested in Burrows & Craig's 2017 Zeta validation (Burrows & Craig 212) using a set of four Shakespeare histories against Marlowe's canon, with an otherwise identical method.

	Number of Segments	Attributed to Author	Attributed to Author %
Doctor Faustus (1604)	49	6	12.2%
Edward II	96	34	35.4%
Jew of Malta	82	19	23.2%
1 Tamburlaine	79	52	65.8%
Dido Queen of Carthage	59	41	69.5%
2 Tamburlaine	81	73	90.1%
Massacre at Paris	42	14	33.3%
Totals for Marlowe	**488**	**239**	**49.0%**
Love's Labour's Lost	96	43	44.8%
Merchant of Venice	97	32	33.0%
King John	94	73	77.7%
Comedy of Errors	64	7	10.9%
Richard III	136	55	40.4%
Taming of the Shrew	95	0	0.0%
1 Henry IV	114	70	61.4%
Totals for Shakespeare	**696**	**280**	**40.2%**

Figure 6: The effect of genre on Zeta. *The Taming of the Shrew* is not recognised as Shakespeare's against a counter set of Marlowe's canon when the base set is four Shakespeare histories. Datasets are size-matched.

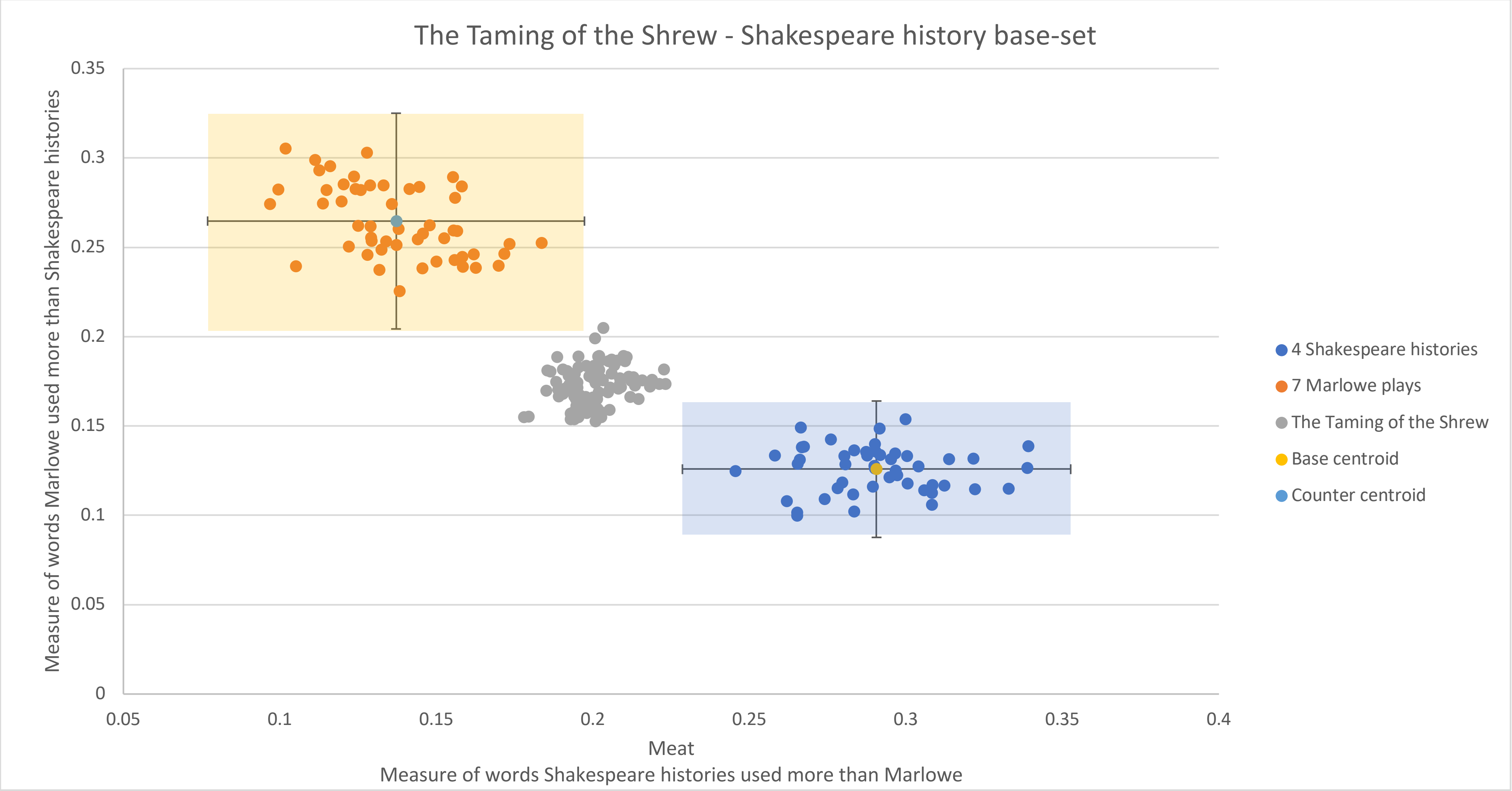


Figure 7: The effect of genre on Zeta. *The Taming of the Shrew* is recognised as Shakespeare's against a counter set of Marlowe's canon when the base set is five Shakespeare comedies. Datasets are size-matched.

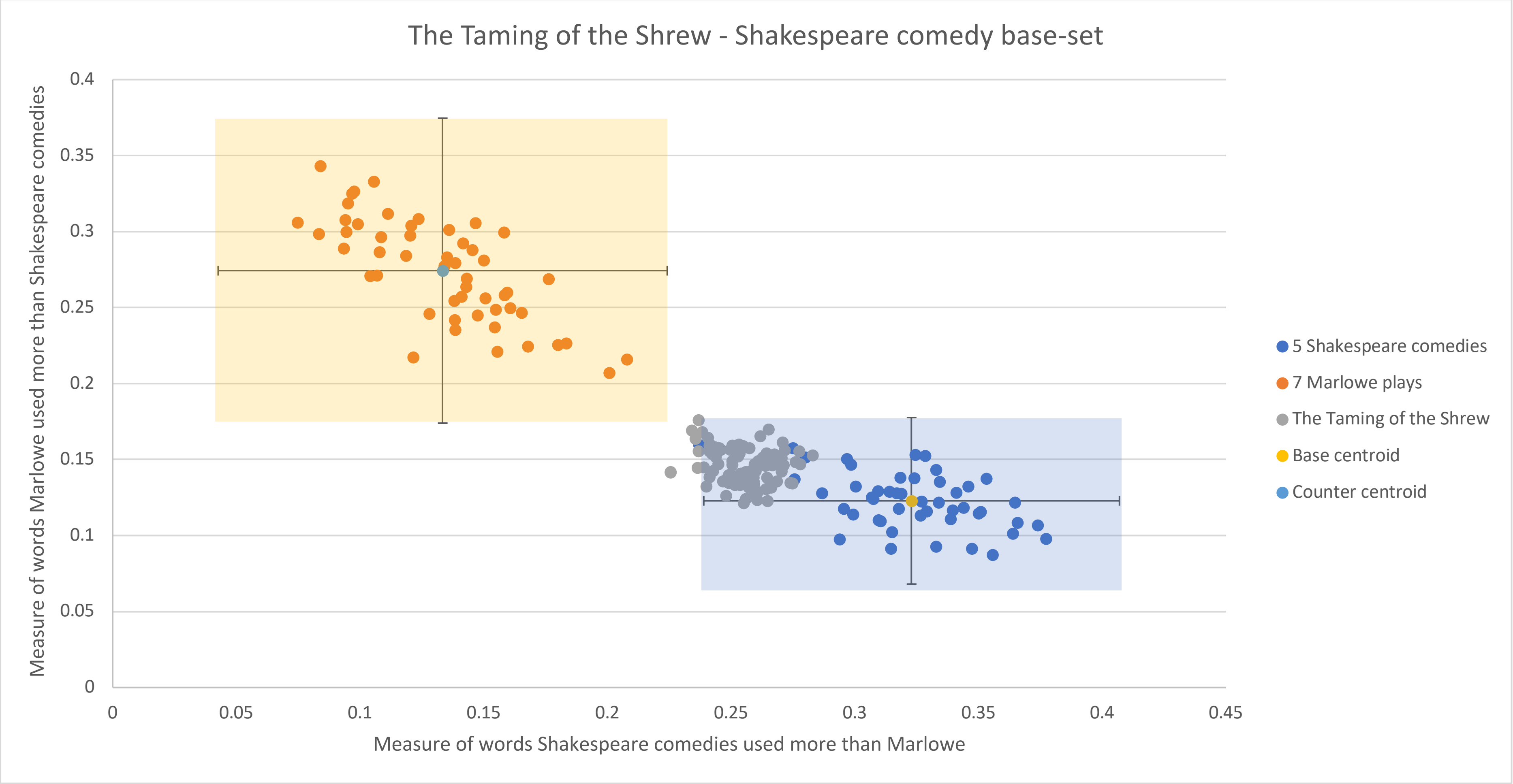


Figure 8. Zeta scatter plot for *3 Henry VI*, using Burrows and Craig's 2017 Marlowe and Shakespeare Zeta validation datasets (Burrows and Craig 212) and the standard deviation attribution method.

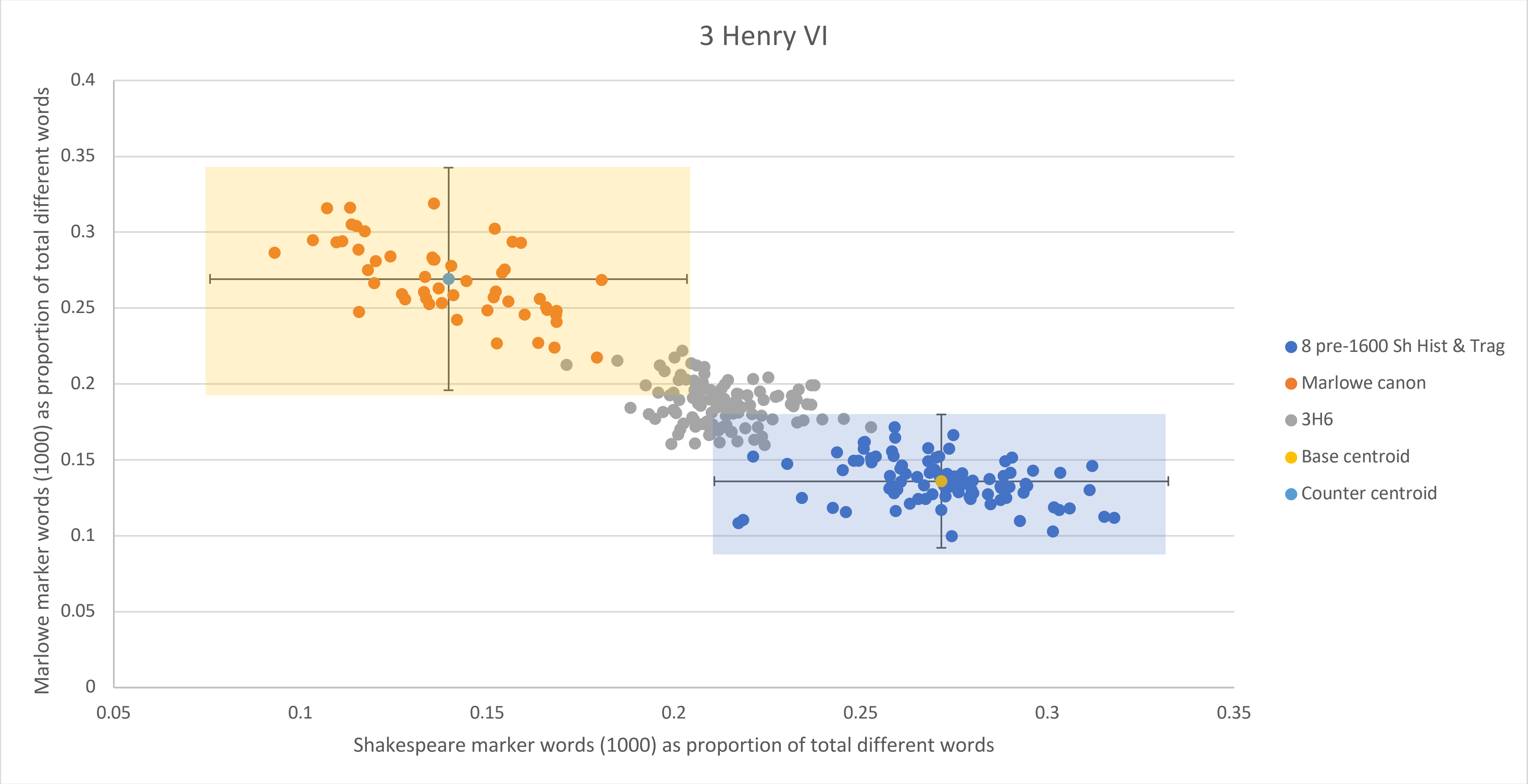


Table 7. Word counts for some of the "top 50" Marlowe marker words that appear in the middle "Joan" sequence of *1 Henry VI* according to Hugh Craig's 2009 Zeta analysis (Craig 63).

Word	Dr F (A)	E II	JoM	MaP	Tam1	Tam2	TOTAL
gold	5	8	**34**	6	**14**	**13**	80
arms	1	**25**	5	0	**35**	**33**	99
realm	1	**25**	0	7	0	2	35
pride	5	8	2	2	7	**12**	36
slain	1	6	9	9	3	3	31
sword	0	**22**	1	6	**20**	**21**	70
golden	3	3	4	1	5	8	24
overthrow	1	3	1	2	**11**	3	21
death	7	**34**	**21**	**34**	**30**	**50**	176
base	2	**16**	6	3	**10**	6	43
damned	7	0	1	2	4	9	23
foe	0	0	1	1	5	9	16
field	0	5	3	3	**12**	**20**	43
yield	0	8	7	0	**13**	**14**	42
cruel	0	4	1	1	5	**10**	21
stay	6	**35**	**18**	7	**13**	**11**	90
conquering	1	1	1	0	**10**	6	19
hell	**42**	7	0	7	**15**	**22**	93
countries	1	1	0	0	0	0	2
words	7	**20**	9	3	**19**	6	64
terror	1	0	0	2	9	8	20

Table 8. The three parts of *Henry VI* as attributed to Marlowe or Shakespeare using the Burrows & Craig's validation datasets and three standard deviations from author centroids.

	No of Segments	Attributed to Shakespeare	Shakespeare %	Attributed to Marlowe	Marlowe %	Not Attributed
1H6	98	14	14.3%	22	22.4%	63.3%
2H6	117	28	23.9%	20	17.1%	59.0%
3H6	111	21	18.9%	10	9.0%	72.1%

Figure 9. Zeta scatter plot for *1 Henry VI*, using Burrows and Craig's 2017 Marlowe and Shakespeare Zeta validation datasets (Burrows and Craig 212) and the standard deviation attribution method.

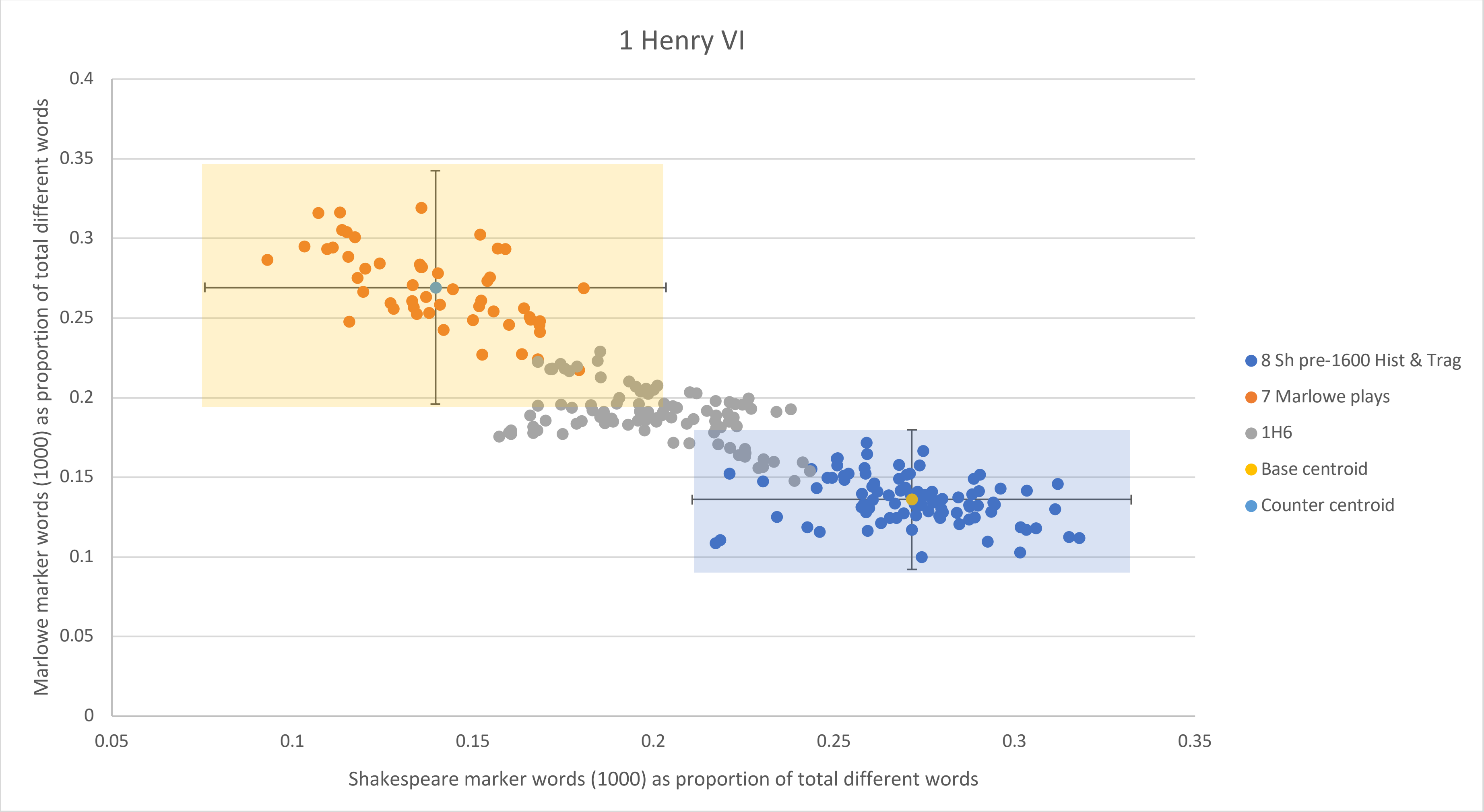


Figure 10. Zeta scatter plot for *2 Henry VI*, using Burrows and Craig's 2017 Marlowe and Shakespeare Zeta validation datasets (Burrows and Craig 212) and the standard deviation attribution method.

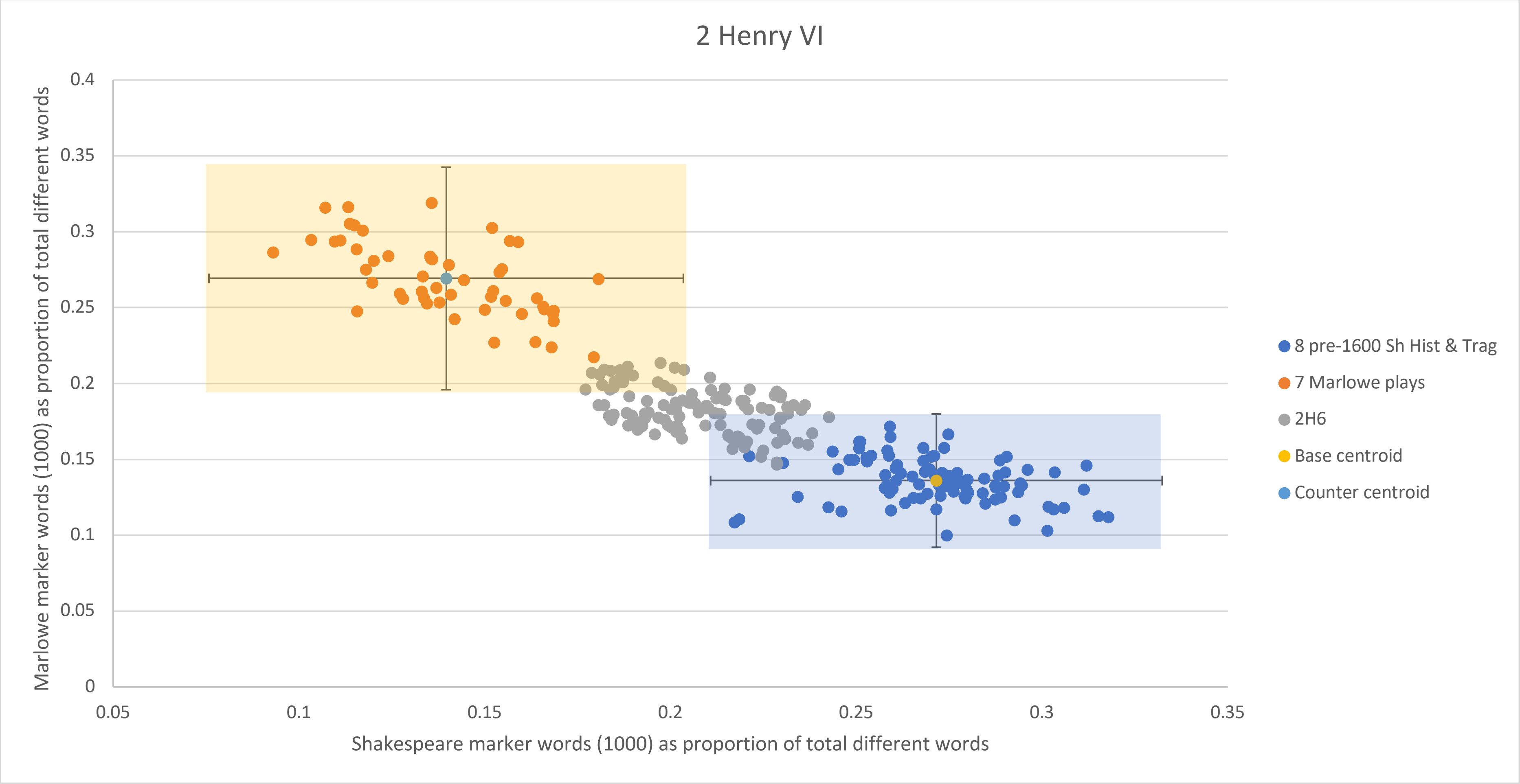
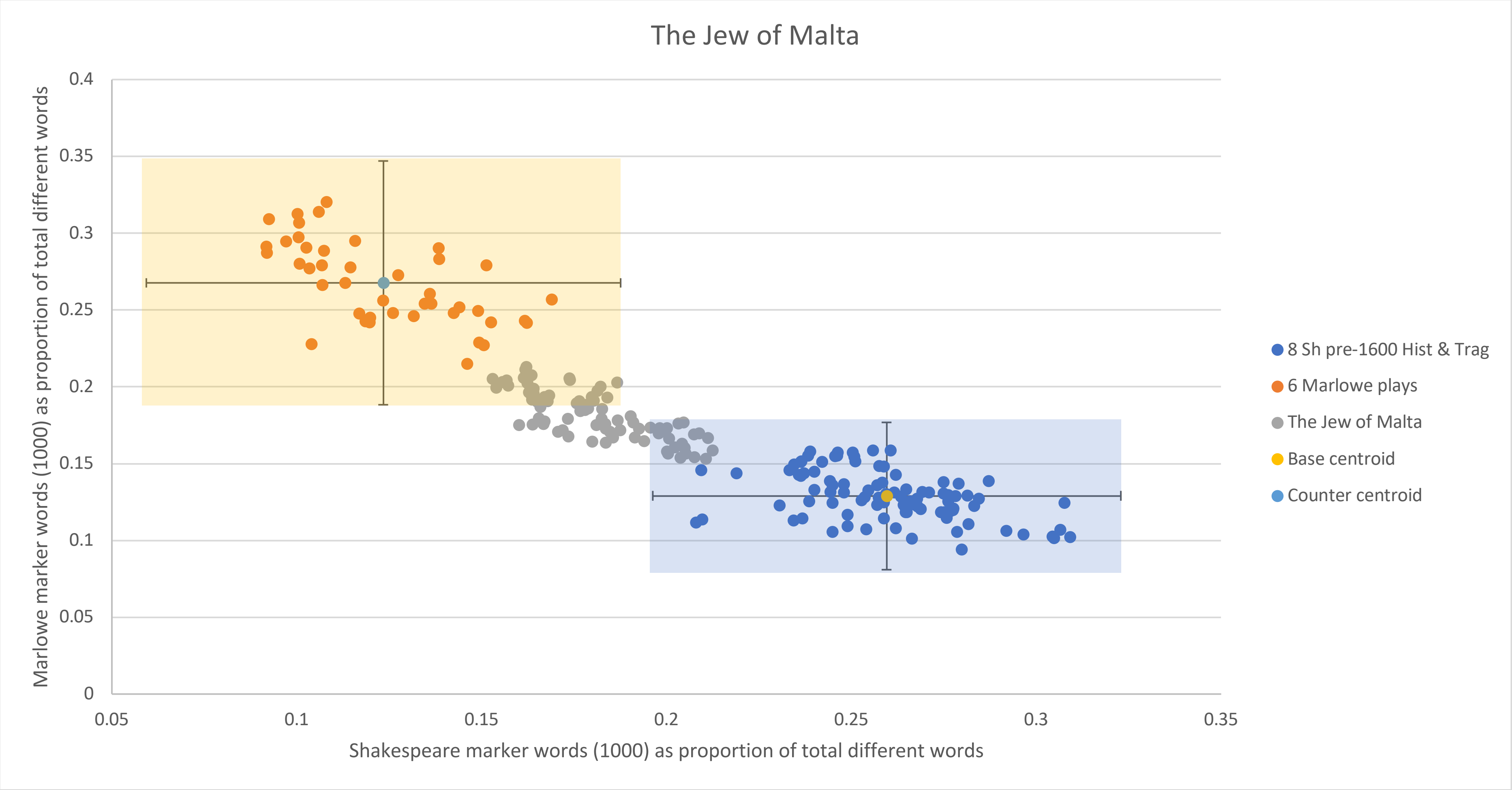


Figure 11. Zeta scatter plot for *The Jew of Malta*, using Burrows and Craig's 2017 Marlowe and Shakespeare Zeta validation datasets (Burrows and Craig 212) and the standard deviation attribution method.



Works Cited

Barber, R. (2019) "Function Word Adjacency Networks and Early Modern Plays." ANQ: A Quarterly Journal of Short Articles, Notes and Reviews, 1-10 DOI: 10.1080/0895769X.2019.1655631.

Barber, R. (2020). BDNE Zeta Dataset. Goldsmiths Research Online. DOI: 10.25602/GOLD.00028390.

Burrows, J. (2007). "All the Way Through: Testing for Authorship in Different Frequency Strata." Literary and Linguistic Computing **22**(1): 27-47.

Burrows, J. and H. Craig (2017). The Joker in the Pack? Marlowe, Kyd, and the Co-authorship of Henry VI, Part 3. The New Oxford Shakespeare Authorship Companion. G. Taylor and G. Egan. Oxford, OUP**:** 194-217.

Craig, H. (2009). The three parts of Henry VI. Shakespeare, Computers, and the Mystery of Authorship. H. Craig and A. F. Kinney. Cambridge, CUP**:** 44-77.

Craig, H. (2017). Authorial Attribution and Shakespearean Variety: Genre, Form and Chronology. Shakespeare Survey 70: Creating Shakespeare. P. Holland. Cambridge, Cambridge University Press. **70:** 154-64.

Craig, H. and J. Burrows (2012). A Collaboration About a Collaboration: The Authorship of King Henry VI, Part Three. Collaborative Research in the Digital Humanities. M. Deegan and W. McCarty, Ashgate**:** 27-65.

Craig, H. and A. F. Kinney (2009). Methods. Shakespeare, Computers, and the Mystery of Authorship. H. Craig and A. F. Kinney. Cambridge, CUP**:** 15-39.

Craig, H. and A. F. Kinney (2009). Shakespeare, Computers, and the Mystery of Authorship. Cambridge, Cambridge University Press.

Eder, M. (2015). "Does Size Matter? Authorship Attribution, Small Samples, Big Problem." Literary and Linguistic Computing **30**(2): 167-82.

Egan, G. (2017). A History of Shakespearean Authorship Attribution. The New Oxford Shakespeare Authorship Companion. G. Taylor, J. Jowett, T. Bouras and G. Egan. Oxford, OUP**:** 27-47.

Elliot, J. and B. Greatley-Hirsch (2017). Arden of Faversham, Shakespearean Authorship, and 'The Print of Many'. The New Oxford Shakespeare Authorship Companion. G. Taylor and G. Egan. Oxford, OUP**:** 139-81.

Hoover, D. L. (2007). "Corpus stylistics, stylometry, and the style of Henry James." Style: 174-203.

Hoover, D. L. (2010). Teasing out Authorship and Style with T-tests and Zeta. Digital Humanities 2010: Conference Abstracts. London, Kings College London**:** 168-70.

Keefer, M. H. (2006). "The A and B Texts of Marlowe's "Doctor Faustus" Revisited." The Papers of the Bibliographical Society of America **100**(2): 227-57.

Logan, R. A. (2007). Shakespeare's Marlowe : the influence of Christopher Marlowe on Shakespeare's artistry. Aldershot, England ; Burlington, VT, Ashgate.

Luyckx, K. and W. Daelemans (2011). "The effect of author set size and data size in authorship attribution." Literary and Linguistic Computing **26**(1): 35-55.

Merriam, T. (2002). "Faustian Joan." Notes and Queries **49**(2): 218-20.

Merriam, T. V. N. and R. A. J. Matthews (1994). "Neural Computation in Stylometry II: an application to the works of Shakespeare and Marlowe." Literary and Linguistic Computing **9**: 1-6.

Rizvi, P. (2019). "The interpretation of Zeta test results." Digital Scholarship in the Humanities **34**(2): 401-18.

Rudman, J. (2016). "Non-Traditional Authorship Attribution Studies of William Shakespeare's Canon: Some Caveats." Journal of Early Modern Studies **5**: 307-28.

Segarra, S., M. Eisen, G. Egan and A. Ribeiro (2016). "Attributing the Authorship of the Henry VI Plays by Word Adjacency." Shakespeare Quarterly **67**(2): 232-56.

Segarra, S., M. Eisen, G. Egan and A. Ribeiro (2019). "A Response to Pervez Rizvi's Critique of the Word Adjacency Method for Authorship Attribution." ANQ: A Quarterly Journal of Short Articles, Notes and Reviews: 1-6.

APPENDIX

The database of all the play texts used in these tests, the code used to run the tests, and instructions for how to run them, can be found at this data repository: https://doi.org/10.25602/GOLD.00028390

Spelling of all texts are standardized if not modernized e.g. 'louyth' has become 'loveth' rather than 'loves'. The Shakespeare plays are the edited Folger Digital Shakespeare texts (https://shakespeare.folger.edu/download/), with errors corrected by Werstine and Mowat. The non-Shakespeare plays from SHC (Shakespeare His Contemporaries), are EEBO-TCP transcriptions with some of their errors corrected by Martin Mueller's students. In the header of the XML files for these plays, Mueller notes that certain common abbreviations and contractions have been expanded. The texts contain the stage directions but exclude the speech prefixes. Character names within speeches remain. Homonyms are not distinguished from one another. A word is defined as whatever the Folger and SHC defined in their XML files as a word, and it is (on the whole) what common sense would lead you to expect. One exception is explored below. Compound words are broken into two separate words, e.g. "kind-hearted" becomes "kind" and "hearted". Verb endings were not altered.

According to the information published in both Craig and Kinney (2009) and Burrows and Craig (2017), they used early printed versions of both Shakespeare and non-Shakespeare plays as copy-texts, standardizing the spelling of function words, and grouping variant spellings of lexical words as single words (thus standardizing spelling throughout, for counting purposes). (2009, xvii-xviii; 2017 198). Some errors clearly remain in the texts (see 'countries' for an example). More detail than this was not available, but in personal correspondence, Hugh Craig confirmed that abbreviations and contractions were expanded, homonyms were not distinguished from one another except for twenty function word 'homographs' marked up for part of speech e.g. whether 'to' was a preposition or an adverb. Compound words were treated as one word. Stage directions and speech prefixes were ignored for the counts. In-text character names were retained. Verb endings were not altered.[17]

The differences between the play texts from the Rizvi database used for these tests, and those used by Craig (whether alone, with Kinney or with Burrows), are therefore as follows:

- All abbreviations expanded (Craig), most abbreviations expanded (Rizvi)
- Twenty function word homonyms distinguished in Craig's texts
- Compound words treated as two words (Rizvi) or one (Craig)
- Stage directions included (Rizvi) or excluded (Craig)

However, there are considerably more similarities than differences between the two datasets: non-Shakespeare plays from the same original texts, spellings standardized but not modernized, speech prefixes removed, in-text character names remain, verb endings not altered. There will be some oddities thrown up by the Rizvi database retaining stage directions: words like 'enter' and exit' might appear higher up the list for Shakespeare's Folio-based texts and the Folger Digital Shakespeare texts than for plays by others which are only minimally edited, thus could end up being marker words in Rizvi-based tests and

[17] Hugh Craig, e-mail message to author, April 30, 2020.

not Craig-based tests. It might also bump character names and descriptors (such as ‘friar’ or ‘nurse’) higher up the list. However, as part of a list of 1000 marker words, I expect these stage direction words to have negligible impact.

To those who object that the higher level of editing of the Shakespeare texts makes them not comparable with the non-Shakespeare texts non-comparable, I would point out that Early modern texts are highly variable in quality. Those in SHC range from bad quartos to good ones possibly proof-read by the author. The kind of consistency we would all prefer in order to run stylometric tests is not possible with Early Modern drama: other researchers in this area (including Craig, Kinney and Burrows) are all using texts of highly variable quality. The difference in errors between edited and un-edited Shakespeare texts is likely to be less than than that between the best and worst non-Shakespeare texts.

There is one oddity that these tests have thrown up, however, which is that the XML texts provided by Folger Digital defines the possessive apostrophe “’s” of the edited Shakespeare texts as a word. Because the possessive apostrophe appears in very few non-Shakespeare texts, “’s” is the top Shakespeare marker in the Zeta tests I have run. This is clearly undesirable. However, all markers hold equal sway no matter what their ranking, according to the current formula for Zeta, so the effect of “’s” is 1/500th or 1/1000th of the overall effect. (I should point out the same is true for “countries” in Craig’s *1 Henry VI* tests).

None of the differences between the two data sources (Craig’s and Rizvi’s) should be an issue; Craig and Kinney clearly state that “[t]he patterns we have uncovered should be robust enough to survive the variations that will arise from using different texts and software, provided the same basic procedures are followed.” (Craig and Kinney Shakespeare, Computers 2009, xviii).