# Function Word Adjacency Networks and Early Modern Plays

**Dr R Barber**

*English and Comparative Literature, Goldsmith of London, London, UK*

Dr R Barber,
ECL,
Goldsmiths, University of London,
8 Lewisham Way,
New Cross
London
SE14 6NW
Tel: 0207 296 4389
E: r.barber@gold.ac.uk

https://orcid.org/0000-0002-1745-9980

Word Count: 7,255 (including endnotes, bibliography, abstract).

# Function Word Adjacency Networks and Early Modern Plays

## Abstract

The Word Adjacency Network method underpinning the *New Oxford Shakespeare*'s attribution of the *Henry VI* plays to Christopher Marlowe as co-author has not been independently tested and is only now being subjected to critiques. The response of Segarra et al. (2019) to criticism by Pervez Rizvi (2018) barely alleviates concerns. This article demonstrates that sections of the plays designated as Shakespeare's were not detected as Shakespeare's by the method according to the authors' own definitions, since his "relative entropy" score was often above zero, which according to Segarra et al. (2016) means the play is no more like Shakespeare's style than it is like the combined style of all six playwrights tested. The disproportionate representation of Shakespeare in the underlying dataset, combined with a mathematical procedure intended to remove "background noise" may explain Shakespeare's hovering around the zero line. A claimed concordance with the results of other stylometric tests giving parts of *1 Henry VI* to Marlowe is demonstrably not present. The high success rates claimed for the method in Eisen at al. (2018) are based on a flawed validation process known as overfitting, an interpretive method altered to improve success percentages, and the effects of disparate canon sizes for which the equations fail to adequately compensate. It is argued that in the light of flaws in the method, and the authors' misrepresentation of their results, the conclusions of both Segarra et al.'s 2016 article and Eisen et al.'s 2018 study should be set aside.

Keywords: Shakespeare; Marlowe: Henry VI; stylometry; computational stylistics; authorship; attribution, WANs, word adjacency networks.

## Function Word Adjacency Networks and Early Modern Plays

The *New Oxford Shakespeare*'s case for Marlowe's co-authorship of the *Henry VI* plays rests on two different types of computational stylistics. Work undertaken with the newest method, based on an analysis of Word Adjacency Networks (WANs) by Segarra et al. was, according to Gary Taylor, the research which "convinced" the editorial board of the attribution (Segarra et al. "Word Adjacency"; Pollack-Pelzner). This method attempts to apply Information Theory to literary language, and "scale up" the work of Claude Shannon, treating a writer's choice of words in a sentence as though it functioned, in probability terms, much like their choice of letters in a word (Shannon). The latter is something which, Shannon showed, has a measure of predictability; as any crossword solver knows, there are only so many letters that can follow the letters SH, for example, and make a valid English word. Within the rules of grammar, a writer's choice of words is considerably larger than the number of letters in the alphabet, but Segarra et al. attempted to reduce this by looking only at a hundred function words, those words that express "grammatical relationships between other words while carrying little or no lexical value of their own" (Segarra et al. "Word Adjacency" 241).[1] The method, which has not been independently tested, proposes that the frequency (and proximity) with which function words are used in relation to each other is an authorial marker.

The decision to analyse function words rather than lexical words supposedly removes the risk that plays will appear similar due to similarities of subject matter rather than style. Lexical word tests are known to be influenced by genre, with historical plays, for example, tending to use similar words (Grieve 255; Burrows and Craig 212). Pervez Rizvi argues that the WAN method makes unproven assumptions about the influences that operate on an author to determine their choice of function words, giving several examples where a function word is determined by the lexical word that precedes it (Rizvi "Critical View" 1-2). Answering this criticism, Segarra et al. state that the validity of their model does not depend on function words being chosen independently from lexical words (Segarra et al. "A Response" 2). But this does not resolve the issue of the necessary independence of function words from what has gone before, which is part of what defines what is known as a Markov chain, and is therefore essential to applying Claude Shannon's theory successfully. In addition, some of the words listed as function words for the purpose of these tests are cause for concern: it's not clear that words like *bar, dare, given, enough* or *might* are really devoid of meaning, and would not be attached to particular characters or dramatic circumstances (Segarra et al. "Word Adjacency" 254).

A paper describing "relative entropy" measured in "centinats" (with mathematics in footnotes) might daunt many Shakespearean scholars, but the authors do what they can to simplify the results by stating clearly that a relative entropy of zero for a play "means that this play is no more or less like that author's work than it is like the combined body of the work of all [the other] authors [in the test]" (Segarra et al. "Word Adjacency" 243). A negative score means it is more like that author's style (according to this test); a positive score that it is less like that author's style. That this is the correct way to read the relative entropies depicted on their graphs is reiterated in the team's 2018 article (Eisen et al. 507).

In practice, rather than looking for a negative relative entropy, Segarra et al. attribute plays and acts “to the author-profile achieving the *lowest* relative entropy” (Segarra et al. "Word Adjacency" 243, my italics). In their 2019 defence of WANs, they say: “the method attributes a text to the author whose profile is closest to the Word Adjacency Network of the text to be attributed” (Segarra et al. "A Response" 5). In fact, the method doesn’t – it is designed such that attributions are only indicated in the case of negative entropy scores. Its inventors have, however, chosen to interpret its results via “lowest wins” rather than “negative scores” attribution.  But if a positive entropy means that an author’s profile is not closer to the text than that of all the other authors combined, as they tell us, then the text should surely not be attributed to them. This “softer” interpretive method has other unexplored implications. With “lowest wins”, Act 1 of *2 Henry VI* is attributed to Marlowe. But if any negative entropy score indicates authorship, then surely Marlowe, Peele and Shakespeare should all be given an authorship credit for this act, rather than just the author with the lowest score?

When interpreting their own results, the researchers note that the figures suggest that “Shakespeare did not write the first act of *1 Henry VI* since it is no more like his profile than it is like Jonson’s profile, and Jonson is an implausible candidate” (Segarra et al. "Word Adjacency" 244). There should be no need to bring in the value judgement (“Jonson is implausible”); the just-above-zero scores for both authors mean, by their earlier definition “it is no more or less like that author’s work” than it is like all tested authors combined.  They claim the similar score of Shakespeare and Jonson “suggest that [Act 1] is by someone we did not profile”, offering Thomas Nashe as a candidate. But another possibility is that the method is at fault.  This is exactly what the same team of researchers conclude when considering the entropy scores for Middleton’s plays (Eisen et al. 509). Finding that Shakespeare and Middleton have near-identical positive scores for *A Game at Chess, Hengist King of Kent*, and *The Revenger's Tragedy*, they do not raise the possibility that these plays were therefore written by an author who wasn't one of the six tested. They assume, rightly, that the method has failed in this case. Shakespeare scores fractionally above zero for Act 2 as well, and no tested author comes under the zero line. By Segarra et al's definition, this means none of the tested authors' styles are "more like" Act 2 than they are like each other, and Act 2 should not be deemed as Shakespearean under this test. This is true not only for Acts 1 and 2 but also Act 5 of *1 Henry VI*, acts 4 and 5 of *2 Henry VI*, and acts 3, 4 and 5 of *3 Henry VI.*

The introduction of "lowest wins" attribution when the method demands "negative scores" attribution is not the only form of misrepresentation present in "Word Adjacency". The waters are further muddied by Segarra et al's scene-by-scene analysis, looking purely at the "relative entropy to profile" of Shakespeare and Marlowe. Act 1 of *1 Henry VI*, suggested in the text as being authored by the untested Nashe, scores in various scenes as either like Shakespeare's style (up to 5 centinats) or Marlowe's (up to 3.5). Analysing scenes in this and other acts, the authors claim that their results "broadly agree with Hugh Craig's determination, using an entirely different method" (Segarra et al. "Word Adjacency" 246). But in fact, they are only in agreement for just under half of the scenes Craig tested, a coin-toss probability.

[Table.1]

Craig did not test every scene in *1 Henry VI*, and some of the scenes in the table were only partially tested; he grouped scenes and part-scenes into early, middle or late sections featuring Joan La Pucelle, so in some respects the results are not truly comparable. However, a glance at Table 1 demonstrates that the two sets of results cannot be described as being in broad agreement.

Another accusation of misrepresentation has already been raised. Pervez Rizvi draws our attention to “a very significant manipulation of the test results” which “has the effect of magnifying differences, perhaps by huge amounts” ("Critical View" 3,4). A number (representing what might be called “background noise”) has been subtracted from all the figures. Rizvi gives the example of a sequence of figures 101, 102, 103, 104 etc. There is a very different relationship between these numbers than that between the same set with 100 subtracted from each: 1, 2, 3, 4 etc. What might be only a very slight difference can be presented as if it is substantial. Segarra et al are right to point out that “investigators often subtract a constant from their experimental results before comparing them” ("A Response" 5) and that this is done “in order to see the distinctions more clearly” ("Word Adjacency" 243). However, their claim that their comparison is absolute (the difference between the two scores) rather than relative (the ratio of one to the other) and therefore cannot be altered by subtracting a constant, is problematic for two reasons.

Firstly, the claim that they are looking at absolute values does not alter the fact that potentially very small differences are being magnified. How accurate are these scores, and what is the confidence interval? The researchers themselves claim an accuracy, for whole play attribution, of 89.6 to 93.6 percent (Segarra et al. "Word Adjacency" 243). But what is the accuracy when looking at fractional differences between relative entropy scores? Let’s take as true, for now, that between 5.4 and 10.4 per cent of the results of this test will be wrong. Consider Act 4 of *2 Henry VI*, which has been given to Shakespeare “with Marlowe being a very close second” (Segarra et al. "Word Adjacency" 246). How significant is the gap between them? To the eye, this looks like the difference between 0.35 and 0.5. Consider that a constant of unknown size has been subtracted to gain these figures. Are we in fact looking at 1.35 and 1.5? Or 100.35 and 100.5? The only clue is Table 3 of “Stylometric Analysis”, which gives relative entropy figures, for six early modern playwrights, in the range 4 to 18  (Eisen et al. 506). If the raw numbers fall within that range – say 13.35 and 13.5 – then they are not significantly different. Is it reasonable for this act to be attributed to Shakespeare when there is no independent confirmation that the method is accurate enough for small differences to be taken seriously?

Secondly, the "absolute not relative" defence ignores the fact that the measure is *supposed* to be relative – to the zero line. In the example just given, both Shakespeare and Marlowe have positive scores which, according to Segarra et al.'s 2016 and Eisen et al.'s 2018 explanations of the method, indicate that neither is the author. Yet even if we interpret results using "negative scores", rather than "lowest wins", we are still left with the problem of what weight to give tiny differences. Should even the smallest showing below zero be taken as a potential indication of authorship? Act 5 of *3 Henry VI* is just on the zero line for Shakespeare. Perhaps its entropy score is -0.01 (we are not told) but how different is this, for our purposes, from zero?

This leads us to the issue of accuracy more generally. The researchers' claim of 89.6 to 93.6 percent accuracy is based on the method's correctly attributing 138 out of 154 early modern plays, and then from a smaller subset whose authorship seems more reliable, 88 out of 94 (Segarra et al. "Word Adjacency" 243). Presumably a small tweak in the method gave a 92.6 percent accuracy rate (87 out of 94) in the later-published study by the same authors (Eisen et al. 510). A typo seems the most likely reason for the claimed top success rate of "94.6 percent" in 2019 (Segarra et al. "A Response" 4). But even the lowest of these figures is inflated for two reasons.

Firstly, as Rizvi points out, Segarra et al.'s claim that the method could attribute 94 sole-authored early modern dramas with 93.6 per cent accuracy is a circular argument; the test was "based on the adjacencies of the one hundred function words … that were found in training to be the most discriminating" on the same set of plays (Segarra et al. "Word

Adjacency" 243; Rizvi "Critical View" 3). Although Segarra et al. defend themselves by saying that they used “leave-one-out cross validation” they have misunderstood Rizvi’s point (Segarra et al. "A Response" 4). The 94 plays used to create a list of function words that were best able to discriminate between those authors were not eligible to be used for a test which supposedly validated the method. Had they used half of those plays to generate the list of discriminating function words they could have used the other half to validate it. But they did not. Using the same dataset for training and validation is a fundamental error and will have inflated the apparent success rate.

The second reason why the claimed success rate is inflated is because it is based on the lowest relative entropy interpretation. If we look purely for negative relative entropies to indicate authorship, which the authors themselves tell us is the correct way to read the results, the success rate drops considerably; only 68 out of 94 plays are correctly attributed, or 72.3 percent.

[Table 2.]

This is not insignificant, being far higher than chance, but it would make WANs, by the authors’ own data, the least accurate of the attribution methods listed in their Table 4 (Eisen et al. 513), lower than PCAs with four principal components, and lower than three different versions of Delta, a function-word test which Craig and Burrows advise is “genre sensitive as well as author sensitive and unsupported results should not be taken either as conclusive or as purely authorial” (Craig and Burrows 36).

Unless compelling reasons can be shown why the “lowest wins” method of interpretation is valid, we should continue to take the zero line as the cut-off point indicating authorship. If we require a negative relative entropy to attribute a section of text to an author, across the fifteen acts of the three *Henry VI* plays, one act would go jointly to Marlowe, Shakespeare and Peele, one jointly to Marlowe and Shakespeare, five would go to Shakespeare alone, and eight acts would be attributed to no-one.[2] If we set a significant negative entropy for attribution at even half a centinat below zero, allowing for a margin of error on the tests, we can see from the graphs that of the fifteen acts, one would be given to Marlowe, one jointly to Marlowe and Peele, and one to Shakespeare. A full twelve acts would not be attributed to anyone.

For all three *Henry VI* plays, analysed by act, Shakespeare hovers around the zero line, at no point scoring more than plus one or less than minus one. We are never given any indication as to what degree of score counts as statistically significant. Shakespeare's hovering around the zero line, rather than being a convincing argument for his involvement, looks more like the effect of his large and varied canon, which as Craig and Burrows noted with respect to a different type of stylometric testing, has a homogenising effect on his test results: "Whatever the target text, it seems, something in Shakespeare is likely to show an affinity for it" (Craig and Burrows 35).

The researchers need to ask themselves a key question. Why does Shakespeare present such a consistent stylistic zero when analysing the *Henry VI* plays by act? A strong possibility is that the subtraction of the constant, which the 2019 defence claims makes no difference to the results, is the reason for Shakespeare's haunting of nil. The subtracted number, which is referred to as the "background reading" for a play, is calculated by comparing a play's WAN score ("entropy") against "the entire set of all the plays by all six dramatists" (Segarra et al. "Word Adjacency" 243). Shakespeare's contribution to the entire set of plays is 28 of the 93, or just over 30 percent.[3] Naturally Shakespeare's plays will show more affinity to this set than the next largest contributor, Jonson (17 percent of the whole, by plays), and certainly more than Marlowe, whose plays comprise only 6.5 percent of the total, and undoubtedly less when relative word counts are taken into consideration. Combine Shakespeare's dominance in the set with the previously noted fact that his canon's size and variability leads to a frequent affinity with other texts, and it becomes obvious that the subtraction of this "background reading" will significantly distort the data.

Other questions arise from considering another of Rizvi's criticisms. Rizvi takes the view that shortcomings in the method's formula led to vital data that might distinguish one author's style from another (relationships between words that always occur in one author's work and never in the other's) being discarded, not for any reason to do with the authorship of texts, but to avoid a mathematically impossible division by zero ("Critical View" 2). Though the authors have attempted to defend this aspect of the method, their counter-arguments are problematic.

Firstly, they claim that because they are analyzing texts of finite length, there is "uncertainty in the observed transitions"; the fact that one author never uses a transition within that finite block does not mean they never use it at all (Segarra et al. "A Response" 3). Their illustrative example is nicely constructed to make their point, since they pick figures that give a ten-fold difference between Authors A and B for the transition *a* to *of* (100 times vs 10 times) and a minimal difference between those authors for the transition *a* to *in* (1 time vs 0 times), a difference which might, hypothetically, be obliterated "if we were to observe a further 1000 words from Author B."  In the example given, it is very clear that the 100/10 difference is far more significant than the 1/0 difference that might "vanish".

But this argument ignores the fact that "uncertainty in the observed transitions" is just as applicable to the transitions they are counting as to those they are discounting.  Consider a situation where a third author, Author C, uses the transition *a* to *in* twice against Author A's once. This transition would be counted. Yet 2 minus 1 has the same value as 1 minus 0, and if we count another 1000 words past the end of the finite-length text the difference between A and C's usage of this transition is just as likely to be obliterated. So why make an exception only for the 1/0 relationship? If "uncertainty" is to be guarded against, in the sense of very minor differences being ignored, then surely a rule should be introduced that ignores all transitions where the differences between authors are minimal and might vanish with further counting i.e. not only 1/0 but also 2/1, 3/2, 4/3 etc. But even this criticism risks losing sight of the fact that the example the authors give does not actually address Rizvi's concern. In making the argument that it is possible that further counting would obliterate any difference, they are ignoring the fact that Rizvi has found 2,786 function word pairs for Marlowe and Shakespeare where that would never happen.

The second defence the authors use is to suggest that Rizvi is restricting himself to what they call "habits of omission" whereas they are looking at "habits of commission", which they refer to as "positive transitions":

> "Of the 10,000 pairs that can be formed from a set of 100 function words … We looked for all such pairings and recorded their strengths for all positive transitions, whereas Rizvi confines himself only to those for which the value in one canon, the k parameter, was equal to zero." (Segarra et al. "A Response" 3)

This misrepresents Rizvi's position. His criticism was that of the 10,000 pairs available, the method looks at less than three-quarters of them, ignoring the 2,768/10,000 (28 percent) of transitions which were used exclusively in one canon and not the other. In defence of Segarra et al., the method measures what it measures. It does not claim to be a complete analysis of function-word usage. Since it relies on chains of words, it cannot measure a non-chain. Where one author doesn't use a particular transition, no chain is created. The restricted scope of the WAN method might be viewed as not dissimilar to the restricted scope of other methods, such as Delta and Zeta, which analyse author styles through specific stylistic lenses.

Nevertheless, Segarra et al.'s hypothetical example is misleading. What it really illustrates is the self-evident truth that big differences (like a ten-fold disparity in usage) are more significant than small ones.  There is no reason to think that in the Marlowe-Shakespeare data, frequencies for both habits of commission and omission are markedly different in nature. Both types of data are likely to show some insignificant differences, and some significant ones. Of the 2,670 words that occur in Shakespeare's canon but not Marlowe's for example, 24 percent of them match the 1/0 of Segarra et al.'s hypothetical example. It is the most common single category. However, 49 percent of the 2,670 function word pairs give a relative usage of 4/0 or higher. Every usage up to 38/0 is represented stepwise, and after that there are some sporadic but notable outliers. The highest usage for a transition that occurs in Shakespeare but not Marlowe is 119/0 for the transition *it* to *one* (although Marlowe's zero is itself nullified when we recognise, as the software I used did, that *tis* stands for "it is").  One of the first pairs I picked at random, the transition of *though* to *nothing*, gave a usage of 13/0. Here are the thirteen instances that occur in the Shakespeare canon:

| | |
|---|---|
| All's Well That Ends Well | Though my estate be fall'n, I was well born, Nothing acquaint… |
| Cymbeline | Though his humour Was nothing but mutation… |
| Henry V | Though all that I can do is nothing worth … |
| Hamlet | Yet, though I distrust, Discomfort you, my lord, it nothing must. |
| Hamlet | Though nothing sure, yet much unhappily |
| King Lear | So your face bids me, though you say nothing. |
| A Midsummer Night's Dream | And though she be but little, she is fierce. HERMIA Little again? Nothing |
| Richard III | Let them have scope; though what they will impart Help nothing else… |
| Twelfth Night | slander in an allowed Fool, though he do nothing but rail; nor no railing in |
| Twelfth Night | a known discreet man, though he do nothing but reprove. |
| Twelfth Night | though she harbours you as her kinsman, she's nothing allied to |
| Troilus and Cressida | Then though my heart's content firm love doth bear, Nothing of that |
| The Winter's Tale | As deep as that, though true. LEONTES Is whispering nothing? |

The fact that Rizvi found 2670 function word pairs that occurred in Shakespeare but not Marlowe, and only 116 that occurred in Marlowe but not Shakespeare, gives us a sense of the effects created by the relative disparities in canon size and variability. Period might also be a factor, both because linguistic fashions change and because individual writers develop stylistically; I note that in the 13/0 example, only *Richard III* was an early play. Out of curiosity, I ran Rizvi's pair-counting code against a reduced Shakespeare canon that was adjusted to match Marlowe's for size and period – *Two Gentlemen of Verona, The Taming of the Shrew, Richard III, Love's Labour's Lost* and *Romeo and Juliet* (113,466 words) against Marlowe's seven plays (113,773 words). Between these two datasets, there are 1,115 function word pairs used by Shakespeare but not Marlowe, and 1,127 function word pairs used by Marlowe but not Shakespeare. In other words, under conditions matched for dataset size and period (but not genre), a 23-fold disparity is reduced to no significant disparity at all. This small experiment tells us nothing about authorship, but it illustrates starkly why disparities in dataset size and period need to be taken into consideration in any stylometric test.

The WAN method, however, does not adequately compensate for dataset size. In "Stylometric Analysis", noting that Marlowe's profile consists of 103,160 words against Shakespeare's 679,256 words, Eisen et al. explain how they use "the size-corrected expression for relative entropy in Equation (7)" (505). This equation adds up relative entropy "only over transitions that are non-zero in every profile being considered" (503). This is, of course, the issue at the heart of Rizvi's "discarded data" critique. In their 2015 paper, where the method was first described, Segarra et al. admitted this calculation was done to avoid division by zero; "to avoid infinite entropies" (Segarra, Eisen and Ribeiro 5467). In the 2018 article it is described, instead, as a form of compensation for vastly different dataset sizes (Eisen et al. 503). The inventors of the method claim that "[t]ransitions rare enough so as not to appear in a profile are, for the most part, also infrequent in all texts" (5467). No evidence was provided to support this claim, but it turns out to be false, as we can see from the 2,670 function word transitions that occur in the Shakespeare canon, but not Marlowe's. Transitions missing from Marlowe’s canon include numerous function word pairs that are commonplace in English such as *though-nothing* (used by Shakespeare, as we’ve seen, 13 times), *can-than* (31 times), *from-an* (35 times), *on-most* (38 times), and *no-nothing* (40 times). When analysing the *Henry VI* plays, the method has discarded 1,368 instances of function word pairs that occur in one canon but not the other, comprising 826 unique function word pairs.

What the “size corrected expression for relative entropy”, Equation 7, means in practice is that when six author profiles are being considered together, as is the case in Eisen et al.’s 2018 study, the number of measured transitions will be even more restricted than the 28 percent restriction that occurs when Marlowe and Shakespeare are compared in isolation. I searched for function word pairs common to all six authors and discovered that only 5,566 out of 10,000 possible function word pairs occur in all six canons (Shakespeare, Marlowe, Jonson, Fletcher, Chapman and Middleton), meaning 45 percent of the possible function word combinations are ignored.

Yet even after discarding nearly half the available data, the effect of mismatched canon sizes are not well-handled by the method. Looking at the function word pairs that occur in every canon, the authors with smaller canons are bound to have fewer of them, even if the proportional usage is the same. Datasets that are, like Shakespeare’s, around seven times larger, clearly contain many more opportunities for those transitions to occur, and potentially to occur with greater proximity, which is the key component measured in the calculation of relative entropy (Segarra, Eisen and Ribeiro 5465; Segarra et al. "Word Adjacency" 237). Thus, despite discarding a considerable proportion of the data, the method will still disadvantage smaller canons and the entropy calculations will be distorted.

The consequence of canon size disparity can be seen clearly in the results of Eisen et al.’s six-author analysis. Their Table 3, which gives relative entropies between profiles, shows Marlowe as having the highest relative entropy against every single author. According to the method, his style is less like Fletcher's, less like Jonson's, less like Middleton's, less like Chapman's and less like Shakespeare's than anyone else's (Eisen et al. 506). One must ask what explains Marlowe being furthest from Shakespeare stylistically, compared with all the other authors, when considerable affinities have been found between Marlowe and Shakespeare's styles, by both traditional and digital humanities scholars, for well over a century (Logan; McDonald 67; Bloom 10; Bakeless 2: 205-67; Brooke xxii; Merriam and Matthews; Merriam). Eisen at al., trusting their method, think it's because the styles, in terms of function word use, genuinely are the most dissimilar. But what is really being measured here is the greatest disparity in canon size. This effect can be seen across the board with Marlowe's results; with the smallest number of words (and thus function word transitions) to analyse, his scores are the highest in every column. Marlowe is the outlier in every subsequent play-by-play test. Eisen et al.'s Table 3 would suggest that the person with the closest style to Shakespeare is Ben Jonson, but what Jonson's low scores actually indicate is that he is the closest to Shakespeare in terms of canon size.

The authors say it themselves: "[t]he more text it has to work with, the more reliable the method" (Segarra et al. "A Response" 1). It is clearly less reliable when it has nearly seven times as much text from one author than it has for another. Believing that Equation 7 has compensated for this, they read the "substantial margin" by which Marlowe is attributed his own plays as being "due in part to the fact that Marlowe's plays were written at least a decade before most of the other authors considered" (Eisen et al. 505). This may be a small factor, as their method has not taken each play's period into account. But what is mostly being illustrated by the notable gap between Marlowe's profile and that of other authors in Eisen et al.'s Figure 5 (511) is actually disparity in canon size.

The other calculation in the WAN method that supposedly compensates for differences in canon size is the so-called "normalization" step of Equation 3. In this step, the researchers “assume that the combined length of the texts written by author [a] is long enough to guarantee a non-zero denominator for a given number of function words” and if this isn’t so, the data is fudged by assuming that a function word is “followed by every other function word in equal proportion” (Eisen et al. 502; Rizvi "A Response to Egan Et Al.") As Rizvi points out on his website, the inventors of the method “make no attempt to explain why this is a reasonable thing to do, when we know even from casual observation that words do not follow other words in equal proportion in any text”. In essence, the procedure pretends that certain function word transitions are present when they are not.

Even if this additional procedure to avoid infinite entropy scores isn’t problematic (though it’s hard to see why it wouldn’t be), Equation 3 also creates the illusion of the unproven and perhaps unprovable assumption at the heart of Word Adjacency Networks. The method’s inventors claim that “it succeeds in large part because it uses a data structure called a Markov chain” (Segarra et al. "A Response" 1), but they have failed to demonstrate that this is the case. The key procedure of Equation 3, aimed at removing the influence of text length, is to divide two numbers by the sum of them both e.g. 2 and 3 become 2/5 (0.4) and 3/5 (0.6). No matter which two starting numbers you choose, doing this will always mean that the sum of the two new numbers is one. In the article that introduced the method, the authors stated that after this procedure, “the normalized networks … can be interpreted as discrete time Markov chains (MC) since the similarities out of every node sum up to 1” (Segarra, Eisen and Ribeiro 5466).

But just because they have introduced an equation where the elements will always sum up to 1, and just because the data “can be interpreted” as a Markov chain, it does not mean it *is* a Markov chain. The *Oxford English Dictionary* defines a Markov process as “any stochastic process for which the probabilities, at any one time, of the different future states depend only on the existing state and not on how that state was arrived at” (OED "Markov, N.") *Stochastic* is defined as “[r]andomly determined; that follows some random probability distribution or pattern, so that its behaviour may be analysed statistically but not predicted precisely” (Oxford English; OED "Stochastic, Adj."). It’s a considerable stretch to see the language of a play, even its function words, as “randomly determined”, and the appearance of some kind of probability distribution has been achieved only by the mathematical sleight of hand of Equation 3. To paraphrase the Markov definition in the context of Word Adjacency Networks, you can certainly process language *as if* it were a Markov chain but to demonstrate that it *is* one, you would have to show that different future function words (i.e. “state” in the *OED* definition) depend only on the existing function word and not on how that function word was arrived at. Considering the complexities of creative composition, it is little wonder that the authors have gone from “can be interpreted as” Markov chains to “uses” Markov chains without the requisite interim step (Segarra, Eisen and Ribeiro 5466; Segarra et al. "A Response" 1).  Markov processes can be used to generate superficially real-looking text, but they have never been shown to generate successful creative works, let alone works of great genius (see Hartman). To analyse the lines of dramatic works (many of them written in iambic pentameter) as if they are Markov chains, when no-one has yet produced so much as a plausible dramatic scene or metrical poem using a Markov process, begs the question. The authors may argue that they are not analysing creative works but rather the function words of those works, which might be assumed to have a more “mechanical” nature. But unless networks of function words in early modern dramas can in some way be demonstrated to *be* Markov chains, this remains an unproven assumption, rather than a strength of the method.

“No matter how it does what it does,” say the proponents of Word Adjacency Networks, “an authorship attribution method deserves scholarly attention if it can be objectively shown to be a good predictor of who wrote what for cases where we already know who wrote what” (Segarra et al. "A Response" 3). But if you overfit your method to make it work for the plays you are testing, it may well achieve apparently high success rates without any assurance that it is accurate in other circumstances or that it is even measuring what you say it is. As this article has elaborated, WANs are only a good predictor of who wrote what when the validation dataset is, against good scientific protocols, the same as the training dataset, and when the interpretative framework demanded by the method ("negative scores") is abandoned in favour of another ("lowest wins") to raise the success rate. The real figures reflect that even with overfitting, it can validate only 43.8 percent of Ben Jonson's plays, and only 30.8% of George Chapman's (see Table 2 above). The only two authors who retain a high (indeed, perfect) success rate with "negative scores" interpretation are Shakespeare and Marlowe, a result easily explained by canon size disparity, since one has the largest canon and one the smallest.

Eisen at al.'s "Stylometric Analysis" gives us the clearest indication that what is being measured with the WAN method in this instance, and mistaken for unique markers of authorship, is the unique canon size of each dramatist. This makes sense because disparities in dataset size mean that the smallest canons, even if they use a function word transition in the same proportion as a larger canon, will not have the same opportunities as larger canons to display a range of proximities which may include closer (therefore higher-scoring) ones.

Nor is the process, as some humanities scholars might assume, a simple crunching of the data. As we've seen, with Equation 3, data that doesn't exist is invented, and with Equation 7, data that does exist is discarded, in both cases to avoid an impossible division by zero leading to infinite entropy scores. In the first instance, the result is bound to be a false reading. In the second, around a quarter to a half of the data is discarded, depending on the numbers of authors being analysed. As Segarra et al. admitted when first describing the method "This is undesirable because the [frequent] appearance of this transition in the text network P1 is a strong indication that this text was not written by the author whose profile network is P2" (Segarra, Eisen and Ribeiro 5467). That might seem intuitively correct but as we've seen, results can be highly skewed by the choice and size of texts.

Both attempts to compensate for different profile sizes have not only been unsuccessful but have caused additional problems or confusions. Where the compensatory process in Equation 7 leads to the discarding of data, the attempted compensation in Equation 3 allows the data to be interpreted as a Markov chain without demonstrating that it is one.

The graphical representation of results with an unknown constant subtracted from every entropy score, though a common scientific practice for ease of reading, would normally be accompanied by the figures themselves, for the sake of transparency. Very small differences in entropy have undoubtedly been magnified, and we are given no indication as to the significance of differences between authors. It seems likely that it is Shakespeare's over-representation in the underlying "all authors" profile (which generates this constant) that is responsible for his entropy scores, across all fifteen acts of the *Henry VI* plays, hovering around the zero line (and more frequently than not, above it, meaning that his profile is not closer to that of the text than the average author).

The Word Adjacency Network method needs to be independently validated under controlled conditions, with validation and training sets separated (i.e. 47 different plays per group), identical profile sizes for each author (e.g. 100,000 words, as in Segarra et al.'s original tests of 2015), a rigid "negative scores" interpretative framework, and significance or margin-of-error calculations included. Segarra et al. should ideally make the process available to other scholars, with open source code so that the steps of the calculation can be fully understood and experimented with.

There are serious question marks hanging over the legitimacy of the Word Adjacency Network method, which Rizvi summarises as being "not suited for use in authorship attribution". If the many issues raised here and in Rizvi's article are given due consideration, the conclusions of Segarra et al. on Marlowe's co-authorship of the *Henry VI* plays, and of Eisen et al. on a range of early modern plays, should be set aside.

## Notes

[1] But this doesn't mean they do not carry meaning. Prepositions certainly carry meaning and a similar argument can be made for pronouns.

[2] As noted previously Act 5 of *3 Henry VI* is just on the line for Shakespeare. We are not told its entropy score but will read it as zero for the purposes of this thought experiment.

[3] There are 93 plays here, as opposed to the 94 of the 2018 article, because they include George Chapman's single combined play *The Conspiracy and Tragedy of Charles Duke of Byron* rather than the separate Conspiracy and Tragedy plays.

## Works Cited

Bakeless, John Edwin. The Tragicall History of Christopher Marlowe. 2 vols. Cambridge: Harvard University Press, 1942.

Bloom, Harold. Christopher Marlowe. Bloom's Major Dramatists. 2nd ed. New York: Chelsea House, 2002.

Brooke, C. F. Tucker, ed. The Shakespeare Apocrypha: Being a Collection of Fourteen Plays Which Have Been Ascribed to Shakespeare. . Oxford: The Clarendon Press, 1908.

Burrows, John, and Hugh Craig. "The Joker in the Pack? Marlowe, Kyd, and the Co-Authorship of Henry Vi, Part 3." The New Oxford Shakespeare Authorship Companion. Eds. Gary Taylor and Gabriel Egan. Oxford: OUP, 2017. 194-217.

Craig, H., and J. Burrows. "A Collaboration About a Collaboration: The Authorship of King Henry Vi, Part Three." Collaborative Research in the Digital Humanities. Eds. Marilyn Deegan and Willard McCarty: Ashgate, 2012. 27-65.

Eisen, Mark, et al. "Stylometric Analysis of Early Modern Period English Plays." Digital Scholarship in the Humanities 33.3 (2018): 500-28.

Grieve, Jack. "Quantitative Authorship Attribution: An Evaluation of Techniques." Literary and Linguistic Computing 22.3 (2007): 251-70.

Hartman, Charles. Virtual Muse: Experiments in Computer Poetry. Hanover, NH: Wesleyan University Press, 1996.

Logan, Robert A. Shakespeare's Marlowe : The Influence of Christopher Marlowe on Shakespeare's Artistry. Aldershot, England ; Burlington, VT: Ashgate, 2007.

McDonald, Russ. " Marlowe and Style." The Cambridge Companion to Christopher Marlowe. Ed. Patrick Cheney. Cambridge: Cambridge University Press, 2004. 55-69.

Merriam, Thomas. "Tamburlaine Stalks In "Henry Vi"." Computers and the Humanities 30.3 (1996): 267-80.

Merriam, Thomas V.N., and Robert A. J. Matthews. "Neural Computation in Stylometry Ii: An Application to the Works of Shakespeare and Marlowe." Literary and Linguistic Computing 9 (1994): 1-6.

OED. "Markov, N." Oxford University Press. Accessed 26 July 2019. <www.oed.com/view/Entry/114201>.

---. "Stochastic, Adj." Oxford University Press. Accessed 26 July 2019. <www.oed.com/view/Entry/190593>.

Oxford English, Dictionary. "Stochastic, Adj." Oxford University Press. Accessed 26 July 2019. <www.oed.com/view/Entry/190593>.

Pollack-Pelzner, Daniel. "The Radical Argument of the New Oxford Shakespeare." New Yorker 19 Feb 2017.

Rizvi, Pervez. "Authorship Attribution for Early Modern Plays Using Function Word Adjacency Networks: A Critical View." ANQ: A Quarterly Journal of Short Articles, Notes and Reviews (2018).

---. "A Response to Egan Et Al." 2019. Date accessed 25 July 2019. <http://www.shakespearestext.com/wan.htm>.

Segarra, Santiago, et al. "Attributing the Authorship of the Henry Vi Plays by Word Adjacency." Shakespeare Quarterly 67.2 (2016): 232-56.

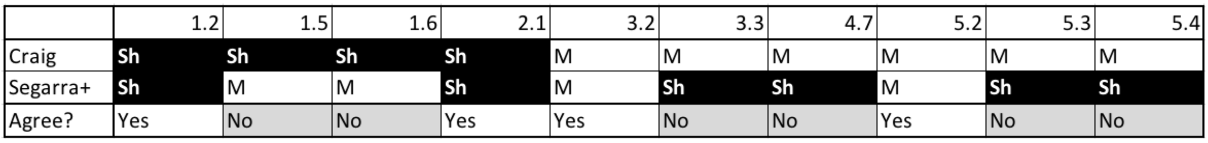
---. "A Response to Pervez Rizvi's Critique of the Word Adjacency Method for Authorship Attribution." ANQ: A Quarterly Journal of Short Articles, Notes and Reviews (2019): 1-6.

Segarra, Santiago, Mark Eisen, and Alejandro Ribeiro. "Authorship Attribution through Function Word Adjacency Networks." IEEE Transactions on Signal Processing 63.20 (2015): 5464-78.

Shannon, C. E. "Prediction and Entropy of Printed English." Bell System Technical Journal 30.1 (1951): 50-64.

Table 1.

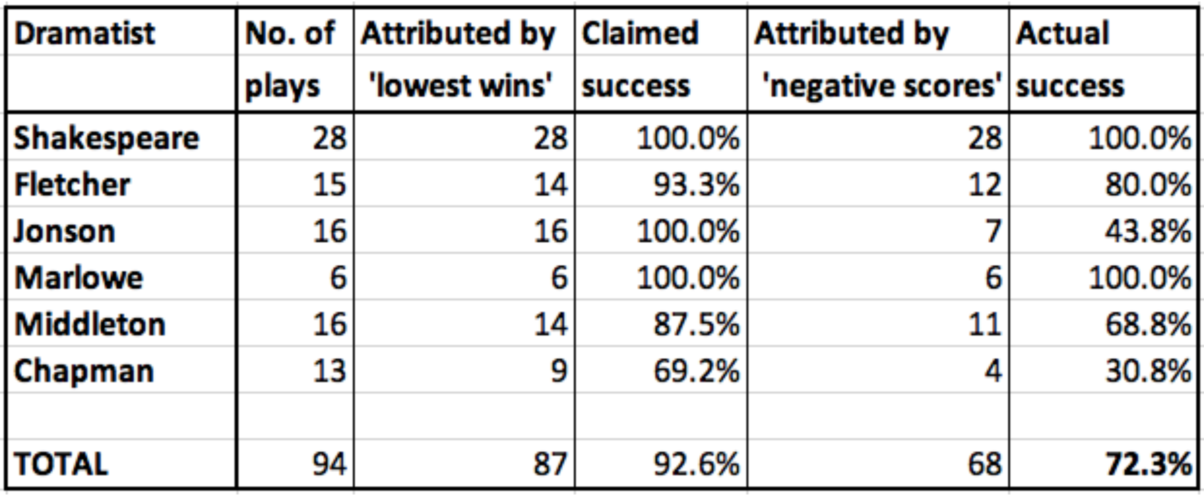
A comparison of the allocation of selected scenes of *1 Henry VI* to Marlowe or Shakespeare by two stylometric methods, Craig's Zeta tests (2009) vs Segarra et al's WANs (2016).



| | 1.2 | 1.5 | 1.6 | 2.1 | 3.2 | 3.3 | 4.7 | 5.2 | 5.3 | 5.4 |
|---|---|---|---|---|---|---|---|---|---|---|
| Craig | Sh | Sh | Sh | Sh | M | M | M | M | M | M |
| Segarra+ | Sh | M | M | Sh | M | Sh | Sh | M | Sh | Sh |
| Agree? | Yes | No | No | Yes | Yes | No | No | Yes | No | No |

Table 2.

Eisen et al's WAN results (2018) with a comparison between the success rates of a "lowest wins" interpretation and a "negative scores" interpretation of relative entropy.



| Dramatist | No. of | Attributed by | Claimed | Attributed by | Actual |
|---|---|---|---|---|---|
| | plays | 'lowest wins' | success | 'negative scores' | success |
| **Shakespeare** | 28 | 28 | 100.0% | 28 | 100.0% |
| **Fletcher** | 15 | 14 | 93.3% | 12 | 80.0% |
| **Jonson** | 16 | 16 | 100.0% | 7 | 43.8% |
| **Marlowe** | 6 | 6 | 100.0% | 6 | 100.0% |
| **Middleton** | 16 | 14 | 87.5% | 11 | 68.8% |
| **Chapman** | 13 | 9 | 69.2% | 4 | 30.8% |
| | | | | | |
| **TOTAL** | 94 | 87 | 92.6% | 68 | **72.3%** |