ANQ: A QUARTERLY JOURNAL OF SHORT ARTICLES, NOTES, AND REVIEWS
https://doi.org/10.1080/0895769X.2019.1590797

A Response to Pervez Rizvi's Critique of the Word Adjacency Method for Authorship Attribution

Santiago Segarra, Mark Eisen, Gabriel Egan, and Alejandro Ribeiro

Pervez Rizvi recently gave a critique of the Word Adjacency Network method for authorship attribution developed by the present authors. The two publications of ours that Rizvi explicitly addresses are Segarra et al. 2016 and Eisen et al. 2018. We thank Rizvi for taking the trouble to consider this method and to look for flaws in it; we believe that such scrutiny by our peers is the best way for the field of authorship attribution to progress. We feel, however, that in this case Rizvi has misunderstood key aspects of the method and misrepresented the things we claim for it. We take this opportunity, therefore, to respond to Rizvi's critique and show where we think he is mistaken about our methods and hence why it still stands as a powerful tool for authorship attribution.

First, a brief overview of the method. As most speakers and writers will have noticed, the commonest words in the English language are short words of the kind *the, and, of, in*, and so on. The equivalent words are also the most common in many other languages. These so-called function words act as a kind of linguistic glue, joining together other words that are more obviously freighted with lexical value, such as *honour, murder, pleasant, passionately*, and thousands of others. (In some languages, the work done by some function words in English is done by alterations to the lexical words by adding prefixes and suffixes and inflecting their endings to represent grammatical relations.) The function words are so heavily used in English that if we take the top 100 of them that occur most frequently across a wide sample of all kinds of writing, these 100 words comprise, by their frequent repetition, about half of everything we say or write.

Investigators have long known that the frequencies of use of particular function words vary from writer to writer in ways that are difficult or impossible for readers to consciously register but that, when counted by automated methods, provide a remarkably distinctive verbal signature of authorship. As we show, their raw frequencies are not the only aspects of function words that turn out to be distinctive. Also distinctive are the habits of putting certain function words near to other function words, so that a writer may prefer to use *and* near to *in* more often than other writers while avoiding putting *the* near to *of*. It seems likely that such habits are largely unconscious, but whether they are or not is irrelevant to the use we can make of them for authorship attribution.

Our Word Adjacency Network method captures the data about the proximities of large numbers of function words, one to another, with a comprehensiveness that no previous method has attempted. It succeeds in large part because it uses a data structure called a Markov chain to hold the data representing each author's preferences and uses methods from information theory (in particular, Claude Shannon's relative entropy) to compare the preferences of one author regarding these proximities with the preferences of another. Applied to the plays of Shakespeare and his contemporaries, we claim a successful attribution rate of 89.6 to 93.6 percent, in the sense that when the Word Adjacency Network method is applied in repeated trials to plays for which the authorship is already known to scholars of early modern drama, the method points to the writer already believed to be the author in 89.6 to 93.6 percent of the cases. The range arises from the method's sensitivity to the size of the sample it is given: the more text it has to work with, the more reliable the method.

Rizvi's first objection is that the Word Adjacency Network method attends only to function words:

CONTACT Gabriel Egan mail@GabrielEgan.com De Montfort University, Leicester, England

> The method takes no account of an author's choice of non-function words, still less their meaning. It regards as indistinguishable the phrases "O, for a muse of fire" (*Henry V*) and "O, for a couple of faggots" (Fletcher and Massinger's *The Little French Lawyer*), since they use the same function words and in the same places. (Rizvi 1)

It is true that the method focuses exclusively on function words. We do not assert that function words are the only places to look for authorship information in a text; different authorship attribution methods seek to recover information on authorial style by focusing on different features of the text. We simply show that there is undeniably useful authorship information to be garnered in the use of function words. Moreover, Rizvi's supposed counterexample is actually evidence in favor of the value of our approach. Fletcher and Massinger, when reproducing Shakespeare's style, kept his exact same use of function words, thereby implicitly accepting the fact that part of Shakespeare's signature use of language is his choice and sequencing of function words. This is true whether Fletcher and Massinger intended to parody Shakespeare or not. Any reader who looks at the sentences would identify them as similar.

What if it is invalid to consider function words in isolation, because the function words chosen by an author are shaped by the lexical words chosen? This is Rizvi's second objection to the Word Adjacency Network method, and to pursue it he undertakes investigations of his own by computational analysis of the same plays we considered:

> The simplest search is to look for cases where the same function word always follows a non-function word. For example, *upside* is listed in the concordance as occurring 23 times, and it is followed every time by *down*. Or consider the word *looker* and its plural *lookers*, which occur 48 times and are followed by *on* every time. Similarly, *devoid* occurs 24 times and is followed every time by *of*. The obvious inference is that the non-function word influenced the choice of the function word that followed it… If it is true, as both our intuition and the textual evidence suggest, that non-function words exercise an influence over the function words that follow them, then the premise of the method I am considering [that is, the Word Adjacency Network method] becomes highly suspect. (Rizvi 1–2)

Rizvi here makes a subtle but fundamental leap in his logical argument. He seems to claim that since our method does not take into account non-function words, we support a view that these words affect neither authorship nor the choice of function words. This is false.

The fact that a given inference method excludes a feature is not an explicit claim that this feature lacks explanatory power or correlation with the included features. To provide an example, imagine that we were trying to estimate the probability of hurricanes occurring in Texas. A successful method might be based on daily values of temperature, pressure, and rainfall. The fact that this method does not include humidity values is not a claim that humidity does not affect the probability of hurricanes or the other included features (it clearly affects rainfall). Satisfactory performance of the method would imply that there is enough information in the considered features to estimate the probability of hurricanes. Going back to our work, non-function words (features excluded from our model) do have information about the authorship (our output) and do have a relationship with function words (which are included features). The exclusion of non-function words does not assume a generative model where authors choose "the non-function word and the function word separately" (Rizvi 2). We neither claim nor assume this at any point, and the validity of our model does not depend on it.

Rizvi's next objection concerns another category of evidence that he thinks we have ignored: the fact that a given author chooses not to follow one of the function words we are interested with another one of those function words. Again, Rizvi's own searches underpin the objection:

> I found 2,670 pairs of function words in which the first word is followed by the second in a Shakespeare play but never in a Marlowe play. For example, Shakespeare follows *about* by *at* (e.g., "they may have their throats about them at that time" in *Henry V*) but Marlowe never does. Conversely, I found 116 pairs of such words in Marlowe but not Shakespeare; for example, Marlowe follows *after* by *any* (e.g., "to imagine that after this life there is any pain" in *Doctor Faustus*), but Shakespeare never does. It is difficult to see how we can have confidence in a method that disregards highly relevant evidence such as this, simply because it does not know what to do with it. (Rizvi 2)

The suggestion that our method "does not know what to do" with this evidence arises because our calculations would assign a value of zero to the occurrence of word *x* followed by word *y* if an author never followed *x* with *y*, and since at a later stage in our method another number would have to be divided by this zero—an operation not allowed in mathematics—we just ignored this evidence, according to Rizvi.

A key point that must be considered here is that we are analyzing texts of finite length. Imagine that in a text of 10,000 words author *A* uses the transition from the word *a* to the word *of* 100 times and the transition from *a* to *in* one time, whereas in another text of the same length author *B* uses the transition from *a* to *of* ten times and does not use the transition from *a* to *in* at all. Which transition encodes better the difference in style? Rizvi suggests that it is obviously the case that the latter transition—the one that occurs in author *A*'s work and not in author *B*'s—is more informative of authorship than the former.

This assumption is not borne out by the evidence. The finite length of the texts introduces uncertainty in the observed transitions. It might well be that if we were to observe a further 1,000 words from author *B*, we would see a transition from *a* to *in* and the difference in style that Rizvi assumes is utterly distinctive would vanish. The transition from *a* to *of* constitutes a quantifiable habit shared (to different degrees) by the two writers. That is, we prefer cases where we have positive data to cases where we do not. Our method encodes this preference by discarding the transitions that appear fewer than k times. For the published method, we selected $k = 1$, so that transitions that do not appear at all are disregarded in the computation of the relative entropy. Of course, it is possible to vary this value along the positive integers as another free parameter and implement known procedures (such as cross-validation) to choose a preferred value for parameter k.

A further consideration regarding what we might call habits of omission (where an author does not follow one word with another) and their relative usefulness for authorship attribution when compared to habits of commission (for which we have positive data) is that Rizvi searches within the Shakespeare and Marlowe canons looking specifically for the latter: word-pairs that are found in one canon and absent from the other. Of the 10,000 pairs that can be formed from a set of 100 function words, he locates 2,670 occurring in Shakespeare's work and not Marlowe's and 116 occurring in Marlowe's work and not Shakespeare's. 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.

No matter how it does what it does, 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. Performed properly, the validation of a method such as ours should be able to quantify how often the method correctly identifies, from its words, the author of a play for which other independent evidence already tells us the author. We claim 89.6 to 93.6 percent accuracy in this regard. Rizvi thinks that we are mistaken about this because we made fundamental errors in our validation. The first error he claims is that:

> The 93.6 percent success rate is claimed for "the ninety-four plays whose authorship is not in dispute." We are told that the method attributed all but six of those plays to their known authors. However, these attributions were obtained after training the method to recognize the correct authors for those very plays. As the article admits: "Each play is attributed … based on the adjacencies of the one hundred function words that were found in training (based on the full, undisputed sole-authored canon of each dramatist) to be the most discriminating" (Segarra et al. 2016 243). The argument is circular. Since the function words used in the tests were changed until they yielded the claimed success rate, the success rate can hardly be used as evidence of the method's correctness. (Rizvi 3)

Here Rizvi has simply misunderstood our method as described in Section IV ("Selection of Function Words and WAN Parameters") of Santiago et al. 2015. At no point in our work does a given play affect its own classification. Rather, the attribution of a given play is affected only by information derived from other plays. This can be mathematically formalized through an argument of statistical independence.

The procedure we choose is a standard one that goes by the name of "leave-one-out cross validation." Imagine we have made a specific choice of function words and want to test the value of this particular choice by measuring its efficacy in authorship attribution. Picture a bag that contains all plays of known authorship and imagine that we select from it a play that we will pretend is of unknown provenance. This play is removed from the bag, and we use the remaining plays to train our classifier, as the jargon has it, which means that we run our method and use it to attribute the play of supposedly unknown provenance. We record whether or not the method identifies as the author of this play the person who we know, from other evidence that scholars agree on, really is the author. We then replace this play into the bag and extract a different one.

We repeat the procedure many times and end up with an accuracy score for this particular choice of function words. The procedure is then repeated for a different choice of function words, and a different score is computed. The choice of function words that performs best is the one that we use to classify plays of unknown authorship. Contrary to Rizvi's assumption (based on misreading our account), the choice of function words is the same for all of the attributions in a test run. Adapting the set of function words to a specific play would indeed artificially increase classification accuracy and would indeed be circular logic. It would also be self-defeating because it would then be impossible to decide what set of words to choose to classify a play that really is of unknown provenance.

A second error that Rizvi identifies in our selection of function words arises because for different experiments we make different selections from the set of 100 function words. As Rizvi notes, our article Eisen et al. (2018) "mentions at one point that it selected 76 out of its 100 function words for use in some experiments, and 55 in some others" (Rizvi 2–3). How can we be sure that we picked the right function words? As Rizvi puts it:

> Now, the number of ways in which we can select several dozen words from a hundred words is astronomically large. The number of ways of selecting 76 words out of 100 is approximately 8 followed by 22 zeroes; and the number of ways of selecting 55 out of 100 is almost a million times greater than that. Therefore, it is no great surprise that if we pick many different subsets from a set of 100 function words, we will eventually find one that gives us the result we are training our method for. Tellingly, despite having literally trillions of subsets of the 100 function words to choose from, the software used in Eisen *et al.* failed to find one that works well for all authors. Instead, it was obliged to use different subsets for different authors. This means that the claimed success rates are of no great significance. (Rizvi 3)

The assertion here is that the possible combinations of function words are so many that we are bound to find one that works well by sheer luck. This is simply not true. We follow a "greedy" approach where we first rank the words (in terms of frequency) and then choose how many words to include. In this way, we first reduce the complexity of the method to be given by the number of function words instead of the set of function words. Second, the complexity of a classifier offers no guarantee that it will work well. In fact, quite the contrary is true, and a lot of effort in machine learning is devoted to finding classifiers of low complexity.

Perhaps this is best illustrated with a counterexample. Let us propose a clearly bogus classification method in which the authorship of a play is determined by the number of times each of 100 function words appears on the first page. There is an innumerable number of possible choices for these 100 words, but none of them would yield a reasonable classification accuracy. We should also point out that while the choice of function words does have an effect on classification accuracy, the effect is noticeable but not dramatic. We can still classify reasonably well with alternative choices of function words. These tradeoffs are described and analyzed in Santiago et al. 2015.

Most importantly, Rizvi misunderstands us when he writes, "Since the function words used in the tests were changed until they yielded the claimed success rate, the success rate can hardly be used as evidence of the method's correctness" (Rizvi 3). No, all of the attributions on which our claimed success rates of 89.6 to 94.6 percent are founded were made using the same set of function words. As we wrote, making this clear, "Each play is attributed to the author-profile achieving the lowest relative entropy, based on the adjacencies of the one hundred function words listed in Appendix 1(b)

that were found in training (based on the full, undisputed sole-authored canon of each dramatist) to be the most discriminating" (Segarra et al. 243).

Rizvi's final objection is that we manipulated our results to make them look more significant than they truly are. This concerns the supposed misrepresentation that occurs when we subtract a constant from each of a series of experimental results. We wrote:

> In order to see the distinctions more clearly, we first calculate for each play its relative entropy with the entire set of all the plays by all six dramatists in order to get a kind of background reading of just how far this particular play differs from the collective norm. Each time that play's relative entropy with the canon of one of the six dramatists is calculated, we deduct from that relative entropy the background reading for that play. (Segarra et al. 243)

To try to show how this distorts the results, Rizvi (3–4) asks the reader to consider the difference between a student who gets 101 on a test and a student who gets 105. If we take 100 away from each result before announcing it, the figures become 1 and 5, and the second student seems to do much better than the first. By contrast, in the original data of 101 and 105, the second student seemed to do only marginally better than the first.

There are two points that are worth mentioning with regard to this comment. The first one is that investigators often subtract a constant from their experimental results before comparing them. Studies of the health risks arising from the explosion at the Chernobyl nuclear power station necessarily subtract the naturally occurring background radiation that we all receive anyway in order to make sense of the increase caused by the explosion. In such cases, the number we are concerned with is not the total but rather the difference between the reading before and the reading after the phenomenon we are interested in.

Naturally, the constant being deducted has to be a reasonable one, not one chosen simply to exaggerate the results. Rizvi's own example of student grades may serve to illustrate that such deductions can be reasonable. Consider the scores for a multiple-choice test in which the candidate must, for each question, select the right answer from a menu of five suggestions. Someone choosing answers randomly will get one fifth (20 percent) of the answers right by chance, so it would be reasonable to discount the first 20 percentage points before comparing students' grades. Consider two students who score 20 percent and 60 percent. On these raw numbers the second student seems to have done three times as well as the first. But if we discount the first 20 percentage points because they represent the background reading, the value that a person would get anyway even without knowing any of the answers, we arrive at a comparison of 0 with 40. This reveals that the first student demonstrated no knowledge of the subject, doing no better than would a person selecting answers at random. Only after deducting the background reading do we see that the second student's attainment was not three times better than the first student's attainment, but vastly better.

The second point that bears mention is the difference between relative and absolute comparisons. In the test score example that we are discussing, a relative comparison is the ratio 105/100 between the student scores and an absolute comparison is the difference (105 minus 100) between their scores.

A relative comparison can be changed, indeed, dramatically changed, by subtracting a baseline. Absolute comparisons are not. In our case, the method attributes a text to the author whose profile is closest to the Word Adjacency Network of the text to be attributed. This is unchanged by a constant shift of all the distances, as is apparent and clearly stated in our articles. The difference between 101 versus 102 and 1 versus 2 would not affect our attribution results in any way; both the validation accuracies and attribution decisions being made would be the same whether this subtraction was performed or not. Therefore, it is perfectly reasonable to deduct this constant to improve the readability of the results. At no point do we use the relative difference between the distances as a measure of attribution, and none of the correct or incorrect attributions depends on this.

Pervez Rizvi has reached the conclusion that our results should be disregarded. However, all his arguments are based on misunderstandings that we have clarified in this response. We stand by our

articles not subjectively but because upon objective consideration of Pervez Rizvi's arguments we found them without merit. That said, there is a point that is, literally, mentioned in passing: "…we may also note that even the claimed success rate of 93.6 percent would cause dozens of early modern plays to be attributed to the wrong authors." This is a valid criticism and one in which we think an honest intellectual engagement is warranted.

Our analyses cannot be used to conclude irrefutable proof of authorship. In fact, Segarra et al. 2016 and Eisen et al. 2018 go to great lengths to explain the limitations of the technique and that it can be used only as evidence and not irrefutable proof. This, we believe, is a limitation that will be shared by any mathematical analysis. There is not sufficient evidence in the historical record to reach conclusions beyond reasonable doubt. This is where the work of other colleagues and the expertise of the community comes into play. The effort of our team to "assign parts of the Henry VI trilogy from William Shakespeare to Christopher Marlowe" provides evidence that Shakespeare and Marlowe co-wrote this play. But it is only because of the past work of the community as a whole that this can be stated with some degree of certainty. More importantly, there are different methods to be tested and different evidence to be evaluated. This future work may bolster our conclusions, or it may prove them wrong. Whatever the case, it is a discussion that is worth having and one in which we are looking forward to engaging with our community.

Disclosure statement

No potential conflict of interest was reported by the authors.

Works cited

Eisen, Mark, et al. "Stylometric Analysis of Early Modern Period English Plays." *Digital Scholarship in the Humanities*, vol. 33, 2018, pp. 500–28. doi:10.1093/llc/fqx059.

Rizvi, Pervez. "Authorship Attribution for Early Modern Plays Using Function Word Adjacency Networks: A Critical View." *ANQ: American Notes and Queries*, 2018, Advance Access. doi:10.1080/0895769X.2018.1554473.

Segarra, Santiago, et al. "Attributing the Authorship of the Henry VI Plays by Word Adjacency." *Shakespeare Quarterly*, vol. 67, 2016, pp. 232–56. doi:10.1353/shq.2016.0024.

Santiago Segarra, Mark Eisen and Alejandro Ribeiro. "Authorship Attribution Through Function Word Adjacency Networks". *IEEE Transactions on Signal Processing*, vol. 63, no. 20, 2015, pp. 5464–547.