ANQ: A QUARTERLY JOURNAL OF SHORT ARTICLES, NOTES, AND REVIEWS
https://doi.org/10.1080/0895769X.2020.1713714

A Response to Rosalind Barber's Critique of the Word Adjacency Method for Authorship Attribution

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

The Word Adjacency Network (WAN) method of authorship attribution was introduced in a series of papers by the present authors between 2015 and 2018. *ANQ* has published two critiques of the method (Barber; Rizvi) as well as the present authors' response to the first of these (Segarra et al.). This paper is a response to the second critique and will attempt to show that Barber repeats several fallacies from Rizvi's critique that the present authors have already addressed and introduces new ones that suggest confusion about what the method actually entails.

English is predominantly an analytic language (as opposed to a synthetic one) in the strict linguistic sense of mainly expressing the relationships between words via additional helper words – called function words – rather than by using inflection. The function words are the prepositions, articles, conjunctions, pronouns, auxiliary verbs, and so on that have little lexical meaning of their own but join together and show the relationships between the content words in a sentence. So common are function words that a list of the 100 words most frequently used in English would be dominated by them and would comprise about half of all that is said and written in the language.

It is known that the patterns of preferences for particular entries in such a list of 100 most frequent words are distinctive of authorship, so that one writer might use *and* and *the* more frequently than is typical and another writer might use them less frequently. The WAN method shows that we can go beyond mere frequencies of use in order to investigate preferences for 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*; this is the "word adjacency" part of our approach. The WAN method captures these preferences using a data structure called a Markov chain (the "network" part of our approach) and uses Claude Shannon's measure of relative entropy to compare the networks – and hence the preferences – of one author with another.

The differences between networks are demonstrably distinctive of authorship, since the authors are able in blind testing using cases where the authorship is already known to achieve a successful attribution rate of 89.6% to 93.6% on whole plays. The attribution success rate falls off as the text samples become small – individual acts and scenes – but in the opposite direction it soon plateaus as the samples get larger. The method requires a canon of at least a few whole plays that are securely attributed, but once this threshold is reached the results level off and adding further plays to the reference set does not affect the results.

In his critique of the method, Pervez Rizvi objected to the use of function words in authorship attribution on the grounds that these words are not independent of the lexical words around them (Rizvi). He gave no reason why such a dependence might undermine an attempt to use function words for authorship attribution and implied that the inventors of the WANs method simply had not thought that such a dependence might exist. (They had thought of it and do not think it relevant.) Barber rejects the WANs inventors' response to Rizvi (Segarra et al.) on this point, arguing that such a dependence would undermine the use of a Markov chain to model function-word proximities, since the WANs inventors' response "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" (Barber 1).

This assertion indicates a fundamental misunderstanding of what a Markov chain is, since this way of modeling transitions of a system from one state to another does not require independence of the nodes (representing the states that are transitioned between) from anything outside of the nodes ("what has gone before"), and it is unclear what such an independence might even consist of. A classic teaching case for this topic is the transition of weather between sunny, rainy, and snowy days. A Markov chain model of these weather transitions does not assume or require that they are "independent" of outside influences such as precipitation, atmospheric pressure, or humidity levels. The purpose of a Markov chain is to characterize the cumulative effect of such influences as transitions from one outcome or state (a sunny day or a rainy) to another.

Barber rejects some of the 100 or so words used in the WANs method as not function words, in particular, "words like *bar, dare, given, enough* or *might*" (Barber 1). As the WANs inventors have explained (Segarra et al. 241n11), linguists do not agree on just which are the function words and they experimented with alternative lists and found that these make no significant difference to their results. Barber gives no reason for suspecting that the words she objects to somehow bias the test in favor of any particular candidate for authorship of disputed works. Indeed, on reflection, it probably does not even matter that the method is dealing with function words as linguists define them: their essential characteristic is more likely to be the extraordinarily high frequency of the 100 most-used words, since this confers relative invisibility in formal as well as informal English.

Barber believes she has found an arithmetical mistake in the WANs work. The problem is that the investigators at one point claim to be looking for the author for whom the relative entropy between their canon and the play in question is negative but "In practice," writes Barber (1), we attribute the play to the author whose canon has the lowest relative entropy to that play. There is no error here. The truth is that the inventors never utilize the sign of the normalized entropy to make an attribution. In fact, the identification of a 0 value was a choice the inventors made to simplify interpretations. This was, in retrospect, unfortunate since it seems to have caused a great deal of confusion. However, all of the attribution results in the paper are made from pairwise comparisons of a text to a profile and this is independent of what interpretation we may give to a positive or negative number.

To see why Barber's point about negative numbers is irrelevant, let us return to the example from the weather. Suppose that we want to identify any given day in New York for which we know the weather as one that belongs to either the summer or winter season. We could define a profile for a typical summer day as one with daytime temperatures above 20 degree Celsius and a profile for a typical winter day as one with daytime temperatures below 0 degree Celsius. We could then attribute a given day to one of these seasons by comparing its daytime temperature with each of the two profiles and assign it to the one from which it shows the smallest difference.

The location of the 0 point on this scale is irrelevant in this comparison. We could shift our measurements from the Celsius scale to the Kelvin scale in which a single degree difference has the same value but which starts at what is known as absolute zero (−273.15° Celsius). Using the Kelvin scale, our winter profile has daytime temperatures below 273.15 K (= 0° Celsius) and our summer profile has daytime temperatures above 293.15 K (= 20° Celsius). Or else we might record the temperatures in relation to the average annual temperature of 12° Celsius so that the winter threshold is defined as −12° and the summer threshold as +8°. None of these shifts of scale would have any effect on an attribution method that measured the difference between the temperature on a given day and the threshold for each of the two profiles, selecting the profile with the smallest difference. What Barber thinks is our arithmetical error in treating positive values for relative entropy as just as evidentiary as negative ones is in fact intentional because all that matters are the differences.

When analyzing whole plays in relation to authorial profiles, the WANs method as we have applied it selects the profile that has the least difference from the play being tested. This tells us which of those authorial profiles the play most closely resembles. Naturally, this method will not identify an author such as Thomas Nashe or Thomas Kyd for which we do not have a profile because

their securely attributed canon is too small, nor one of the many authors whose works are classified together as simply Anonymous. The WAN method enables us to select between authors of known habits, answering not the question "who wrote this play?" but the question "which of these authors is most likely to have written the play?"

Where there are preexisting reasons to favor two candidates in particular, we can use the WAN method to select between only those two. This is useful in cases where external evidence already points to a pair of collaborating coauthors – as with *The Two Noble Kinsmen* by William Shakespeare and John Fletcher – and we can turn to individual acts and scenes, asking which of the two writers' profiles each of these sub-units most closely resembles. Such a narrowed field of candidates helps mitigate the problem that the method becomes less accurate as the textual samples get smaller. In all such applications, it must be remembered that we are asking "if the author is one of these writers, which is it more likely to be?"

We stress this point about finding the smallest differences in values because Barber's critique is fixated on a false contrast between negative and positive differences in entropy, asserting that only the negative ones are indicative of shared authorship. In our work, no attribution decisions or validation accuracies are based upon the "distance from 0" results, as Barber seems to think. This fundamental misunderstanding takes Barber on a long series of objections to problems that are actually not present in the research that she is discussing. She also fails to notice that procedures described in the validation stage of the method are not the same as – need not be the same as – the procedures used in the application, so that one part of the process may produce differences that span the number zero (some negative, some positive) and others may produce differences that are all positive; in both cases the WANs inventors looked for the smallest ("lowest") differences.

Barber explicitly repeats Rizvi's already refuted (Segarra et al.) claim that subtracting a constant from each result of the experiments exaggerates the outcome. Barber's objection makes no more sense than Rizvi's because she clings to the idea that all the results are relative to zero in the sense that positive numbers show no clear attribution and only negative ones tell us anything. In effect, Barber repeats Rizvi's mistaken objection (Rizvi) to the practice of moving all the data points of a result up or down by the same amount to facilitate easier comprehension of the differences between plays; such a translation up or down the scale does nothing to affect the differences between the scores, which is what is at stake here. We hope that the above analogy with temperature recordings sufficiently illustrates why this is fallacious.

Barber presents a table (her Table 1) comparing the results of the WANs method regarding the play *1 Henry VI* to Hugh Craig's independently derived results for that play. Both investigations concluded that Christopher Marlowe's writing is present in the play, but Barber misrepresents this by tabulating only the scene-based conclusions from Craig and comparing them with the WANs results. The WANs inventors gave a detailed account of where their method agrees and disagrees with Craig's results (Segarra et al. 246–49), and by looking only at "selected scenes" (as her table's caption puts it) Barber gives the false impression that the two studies disagree.

Barber repeats Rizvi's already refuted claim that the validation method used by the WANs team was faulty, asserting that the 94 plays used to find the best words for discriminating between authors should not have also been the 94 plays used to validate the method. According to Barber, the WANs investigators should have set aside fully half the set of 94 plays, using the first 47 plays for finding which function words are most discriminating and the second 47 plays for validating the method (Barber 3). This is a false objection because as the WANs inventors already pointed out (Segarra et al.) it makes no significant difference which function words are used: we could just pick the top 100 most-used words in English and the results would be about the same.

To be more precise, the greedy method used to select the function words reduces the corresponding hyperparameter to a single number as opposed to selecting among all possible subsets of function words. This has the effect of regularizing the method and, thus, avoiding overfitting. In this context, the leave-one-out cross-validation method used by the WANs inventors is perfectly acceptable for assessing the accuracy of the method, since the constitution of the function-word list does not

substantively affect the results and at no point is the play being attributed used in the constitution of the corresponding profile WAN of its true author.

Barber repeats Rizvi's objection that the WANs method does not count those transitions (from one function word to another within five words of it) that are entirely absent in the plays under consideration (Barber 5–6). This objection has already been answered (Segarra et al.) with two rebuttals. The first is that there are many features of language that the WANs method does not count and that itself is not a valid critique since every method counts only some of the many things we might count. The second is that counting absent transitions merely privileges the threshold of zero occurrences, making a purely binary determination (zero or non-zero?) whereas the WANs methods instead use the richer data arising when we count whether things happen once, twice, thrice, and so on.

Someone could build a classifier based on the absence of function-word proximities, as Rizvi and Barber prefer, and we could then compare this classifier's power with that of the WANs method to see which is best. But simply pointing out that such another classifier might be built, as Barber does at length, is not of itself a critique of the WANs classifier. Sketching what such a classifier might attend to, Barber lists the function-word proximities for *though* to *nothing* that occur 13 times in Shakespeare's plays and not at all in Marlowe's, and it is clear that her criteria are different from those of the WANs method: she admits proximities much greater than the five-word window of the WANs method, and she admits proximities that span a change of speaker, as the WANs method does not (Barber 5–6).

Using the function-word transitions that occur in Shakespeare but do not occur in Marlowe, Barber finds that canon size matters in the sense that the smaller Marlowe canon has less opportunity, as it were, to use many of the transitions found in the much larger Shakespeare canon. Indeed, this dependence on canon size is a good reason not to set the threshold for counting at zero but instead to look at those transitions that Shakespeare and Marlowe both use but to differing degrees, as the WANs method does. Barber then attributes to the WANs method the fault that applies only to her own method, insisting that the problem of canon size "illustrates starkly why disparities in dataset size and period need to be taken into consideration in any stylometric test" (Barber 6).

Barber objects to the WANs inventors' claim that transitions that are so rare that they occur not at all in some writers' canons are in fact very rare in all writing, offering as counter-evidence the many transitions she found in Shakespeare's canon but not Marlowe's. But Barber's counter-evidence transitions are numerous because of her broader filter that admits words further apart (even spanning two speeches) than the filter of the WANs method, as well as the relatively large size of Shakespeare's canon. Moreover, as already mentioned, the fact that there might be some stylistic markers overlooked by the WANs method does not constitute a valid argument against the method, since this argument can be made against any authorship attribution procedure.

Barber thinks that canon size affects the WANs method even though this method counts only transitions that occur in the play being tested and the canon in question, and this is true but only up to a point. As the WANs inventors explained, and indeed repeatedly quantified, the accuracy of the method falls off as the texts being examined become small; hence, the reliability of the method is poorer for individual scenes from a play than for whole acts or whole plays. But once a certain minimal canon size is met it is not at all true, as Barber claims it is, that authors with large canons are treated differently from authors with small canons. Above a certain threshold the size of the texts makes no difference because the method measures not the frequencies of occurrences of the function words but their relative proximities one to another and these stop changing significantly once the texts are above a certain size.

Barber thinks that a table published by the WANs inventors (Eisen et al. Table 3) actually shows the effect she claims, but she is misreading the table. It does not show "Marlowe being furthest from Shakespeare stylistically, compared with all the other authors" (Barber 7) since their relative entropies are 8.9 and 10.1 – there are two numbers since it matters which author we put first and

which second in this non-commutative calculation – and other distances in the table are greater. Looking along the table's Shakespeare row (that is, showing "from Shakespeare stylistically") the Fletcher distance is 8.9, same as Marlowe's, and looking along the Marlowe row the Fletcher distance is 17.4, which is greater than the Marlowe-Shakespeare distance of 10.1.

Whichever way we parse Barber's claim, the table simply does not support it. According to Barber, "what is really being measured here is the greatest disparity in canon size" (Barber 7) but the table does not bear this out either, since the Chapman and Fletcher canons used are about the same size (13 and 15 plays, respectively) and yield distances of 9.6 and 8.4, while the roughly equally sized canons of Chapman and Jonson (13 and 16 plays) yield distances of 5.8 and 5.4. Put another way, Chapman's canon of 13 plays being less than half the size of Shakespeare's at 28 plays (referring as ever to the plays tested) does not have the effect that Barber claims follows from a "disparity in canon size," since their distances are 4.7 and 4.8, the lowest numbers in the table. The WAN method is measuring real stylistic differences, not canon sizes, as indeed is already clear from the multiple validation runs that show the method having better success in blind attributions (that is, when making attributions for plays for which we already know the answer) than other methods in use.

Barber quotes from the WANs inventors an explanation of one of their equations: they "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." Barber does not appear to understand the process of normalization that this equation performs and she quotes Rizvi's mistaken belief, offered in an unpublished paper, that the WANs inventors fudge their data by assuming that a function word is "followed by every other function word in equal proportion." Rather, because the WANs method is concerned with the differences between the frequencies of transitions, the normalization step perfectly reasonably records that there is no measurable difference when there is no transition to record. This is a commonly used strategy to avoid absorbing states in classification or ranking methods based on Markov chains, such as the celebrated PageRank algorithm to sort webpages. In any case, as also explained by the authors, the appearance of function words with a "zero denominator" is rare for the set of words chosen and, hence, the strategy chosen for the normalization of this rare occurrence ends up having a negligible effect on the ultimate classification.

Toward the ends of Barber's critique is becomes apparent that she does not understand the mathematical system, the Markov chain, whose application she is objecting to. Rather than accepting that Markov chains are a way of looking at certain phenomena, Barber believes that some phenomena actually are Markov chains and others are not: "… just because the data 'can be interpreted' as a Markov chain, it does not mean it *is* a Markov chain" (Barber 8, emphasis in original). This is akin to claiming that just because the sum of the squared differences from their mean that is shown by a series of numbers can be understood as their variance this does not mean that this sum *is* their variance. That is, Barber is taking a mathematical method for making sense of the world and mistreating it as an assertion about the nature of reality.

Instead of using a reliable introduction to the topic of Markov chains, Barber turns to the definition in the *Oxford English Dictionary* and finds that it relies on the notion of a "stochastic" process, which in turn she looks up and finds that it concerns random probability distributions. For Barber, this reveals the weakness of the whole approach since "It's a considerable stretch to see the language of a play, even its function words, as 'randomly determined' …" (Barber 8). In Information Theory, it is widely accepted that language generation can be studied as a stochastic process that may be modeled by probability distributions, and indeed the practical successes of such services as Google's Translate tool and the impressive language-generation system GPT2 prove that this modeling works. Put another way, the Markov chain is a model to study function-word occurrence in Early Modern English writing; it is not a generative process. The legitimacy of this model is established through the validation process which demonstrates that it can differentiate writing styles in known cases. But it takes more than the *Oxford English Dictionary* entries for "Markov process" and "stochastic" – Barber's only stated sources – to make sense of work in this field.

In the summary of her essay, Barber claims that the WANs method overfits to its training data. She offers no evidence for this claim, and if it were true the method would not achieve the high success rate that it shows in the validation runs, as extensively studied in (Segarra et al.). Barber, of course, rejects these validation runs too and insists that all the method can do is attribute plays by canon size and hence Marlowe stands out because his canon is so small. Even without the validation runs, however, Barber's conclusion is clearly contradicted by the non-Marlovian results. The WANs inventors' experiments confirm Shakespeare's coauthors George Peele in *Titus Andronicus*, Thomas Middleton in *Timon of Athens*, and John Fletcher in *Henry VIII* and *The Two Noble Kinsmen* (Segarra et al. 249–51; Eisen et al. 518–521, 524–525). Barber is silent on these results, which cannot be explained by relative canon sizes.

Disclosure statement

No potential conflict of interest was reported by the authors.

Works cited

Barber, Rosalind. "Function Word Adjacency Networks and Early Modern Plays." Advance access. *American Notes and Queries*, 2019. doi: 10.1080/0895769X.2019.1655631

Eisen, Mark, et al. "Stylometric Analysis of Early Modern English Plays." *Digital Scholarship in the Humanities*, vol. 33, 2018, pp. 500–28.

Rizvi, Pervez. "Authorship Attribution for Early Modern Plays Using Function Word Adjacency Networks: A Critical View." Advance Access. *American Notes and Queries*, vol. 2019, 2018. 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.

Segarra, Santiago, et al. "A Response to Pervez Rizvi's Critique of the Word Adjacency Method for Authorship Attribution." *American Notes and Queries*, 2019, TBA. doi: 10.1080/0895769X.2019.1590797

Segarra, Santiago, et al. "Authorship Attribution Through Function Word Adjacency Networks." *IEEE Transactions on Signal Processing*, vol. 63, 2015, pp. 5464–78. doi: 10.1109/TSP.2015.2451111.