# Summary

The purpose of this article is to show one way in which some of the data I have published online in 2017 and 2018, in my [Collocations and N-grams](http://www.shakespearestext.com/can/) site, can be used. I do this by attempting to answer an important question in authorship studies.

Are short N-grams, such as 1-grams and 2-grams, better or worse for authorship attribution than longer ones such as 4-grams and 5-grams?

Until recently, researchers had been led astray by the Antonia et al article, which purported to show that short N-grams, as well as function word skip N-grams, are more reliable authorial markers than longer N-grams. That article has now been discredited (Rizvi 2018) and the question I have posed above once again seeks an answer.

The results below suggest that, at least as far as this test is concerned, 4-grams are the best authorial markers.

# Test Method

The tests I have done are based entirely on the formal N-gram counts I published online in June 2018. They may be reproduced by any researcher, either using her own choice of software or by using the appropriate Attribution Tester spreadsheet,[[1]](#footnote-1) which I also provided.

I created a notional ‘walled garden’ of plays for these tests, considering N-gram matches only between those plays, disregarding matches with all other plays. The plays I chose are the ones listed by Segarra et al, as they are held to be sole-authored plays whose attributions are not in doubt. My purpose is to test how accurately the counts I have published can attribute those plays to their known authors.

The Segarra et al list contains 94 plays. My database contains 86 of them; and those 86 are the ones used in all my tests.[[2]](#footnote-2) Those plays are by six authors: Chapman, Fletcher, Jonson, Marlowe, Middleton, and Shakespeare.

The test method is as follows.[[3]](#footnote-3) I have performed the test separately for (i) all N-grams, (ii) unique N-grams, and (iii) function word skip N-grams. For simplicity, the description below is for all N-grams.

For each of the 86 plays, and for each value of N from 1 to 10, find out how many formal N-gram matches there are between it and every one of the other 85 plays. That is 850 numbers per play. For example, for *Hamlet*, these 850 numbers are found in the file **summary-formal-ngrams-tokens-hamlet.csv**.

Total the information collected above for each play, at author level. This means that we obtain 5,160 totals (because 86 plays times 6 authors times 10 = 5160). In other words, for each combination of play and author, and for each value of N from 1 to 10, we note how many N-grams that play shares with plays by that author.[[4]](#footnote-4)

Each total must now be weighted, to avoid bias in favour of authors with large canons, such as Shakespeare. Formally, if P is a play and A is an author, then our notation is:

The weighted total WN is then defined as:

To avoid dealing with tiny numbers with lots of decimal places, I multiplied each WN value by 10,000,000 and rounded to the nearest integer. This makes the data more presentable, without changing the outcome of the tests.

For each play and each value of N, we now have six WN values, one per author. We may now define our test outcome:

We say the outcome is Green if the WN value for the known author of the play is higher than each of the other five WN values. This means our test has correctly attributed the play.

We say the outcome is Red if the WN value for the known author of the play is less than at least one of the other five WN values. This means our test has attributed the play to the wrong author.

We say the outcome is Amber if the WN value for the known author of the play is not less than any other WN value, but one or more other WN values is equal to it; in other words, a tie for first place. This means our test was inconclusive. It typically happens when there is little or no data; for example, if there are no N-gram matches at all for the value of N we are using.

With the above rule, we obtain ten outcomes for each play, one for each value of N. In the sections below, I give the raw data; that is, the 60 WN values per play. I also give the ten outcomes per play, one for each value of N from 1 to 10, so we can see which values of N provide the most Green outcomes.

# Test Results – All N-Grams

The raw data is given in this spreadsheet:



The test outcomes are in this spreadsheet:



We see that 4-grams correctly attribute 82 out of 86 plays, while 5-grams are not far behind, with 80 out of 86 correct.[[5]](#footnote-5) Other values of N are much less accurate.

With 4-grams, three out of the four failures are for Chapman, while the fourth is for *The Jew of Malta*, a play that other methods also find hard to attribute correctly. *The Faithful Shepherdess*, which other methods struggle with, is correctly attributed, as are all plays by Fletcher, Jonson, Middleton and Shakespeare. There is clearly a difference between Chapman’s works and those of the other five authors, as far as what this test measures is concerned, which a Chapman scholar may be able to explain.

# Test Results – Unique N-Grams

My published counts are for matches which are unique across the whole database, and those are the counts I have used for this test, rather than calculate freshly the N-gram matches that are unique just within the 86 plays used in the test.

The raw data is given in this spreadsheet:



The test outcomes are in this spreadsheet:



We see that unique 3-grams get 84 out of 86 attributions correct, with unique 4-grams only just behind at 83 out of 86 correct. *The Jew of Malta* gets correctly attributed by both. Other N-grams are much less accurate.

The two failures by unique 3-grams are a play by Chapman and, more interestingly, *The Taming of the Shrew*, which may be a co-authored play after all, as has long been suspected.

# Test Results – Function Word Skip N-Grams

The raw data is given in this spreadsheet:



The test outcomes are in this spreadsheet:



These are bad results. We see that 5-grams are the least bad, but even they only get 63 out of 86 correct. The rest are even worse. At least as far as this test is concerned, function work skip N-grams are completely useless.

# What about Types?

The tests I have done use counts of matching tokens, not types. As I have written [elsewhere](http://www.shakespearestext.com/can/counting.htm), authorship attribution tests using types are to be avoided if other things are equal, since they are much harder to do, it being not possible to add counts based on types. In a test using types, every number must be separately worked out. Token-based methods are scalable in a way that type-based methods are not.

I may come back, and present results calculated from types. For now, suffice it to say that informal tests have convinced me that types produce less accurate results from this test than tokens.

Pervez Rizvi

October 2018

WORKS CITED

**Antonia, A., Craig, H. and Elliott, J.** (2014). Language chunking, data sparseness, and the value of a long marker list: explorations with word n-grams and authorial attribution, *Literary and Linguistic Computing*, 29: 147-163.

**Rizvi, P.** (2018). [The Interpretation of Zeta Test Results](https://academic.oup.com/dsh/advance-article/doi/10.1093/llc/fqy038/5078547?guestAccessKey=a9c3b348-9388-4c9d-a591-922fd2506f39), *Digital Scholarship in the Humanities*, print publication forthcoming.

**Segarra, S., Eisen, M., Egan, G. and Ribeiro, A.** (2016). Attributing the Authorship of the *Henry VI* Plays by Word Adjacency, *Shakespeare Quarterly*, 67: 232-256.

1. If you use the Attribution Tester then note that it attributes *The Revenger’s Tragedy* to “Middleton, Thomas (?)” rather than to “Middleton, Thomas”, since I took the attributions from the now-defunct *Shakespeare His Contemporaries* website. For these tests, I removed the question mark. [↑](#footnote-ref-1)
2. My database divides *King Lear* into two divisions: Folio-only lines, and the rest of the play. For convenience, I used the latter division in these tests rather than merge the two divisions back into one. [↑](#footnote-ref-2)
3. We may distinguish between two types of attribution methods, **stable** and **unstable**. A stable method is one which is defined once and is then always the same no matter which texts you use it for. Zeta, for example, is largely stable – perhaps we may say semi-stable – but not wholly so. Having chosen the texts, the researcher must still choose which of them are in the base and which in the counter, and how many types to use when calculating the Zeta scores. Nevertheless, there is a high degree of objectivity, and therefore stability, in Zeta. Function word adjacency network (Segarra et al 2016) is an extreme example of an unstable method. The researcher must choose one from literally trillions of possible sets of function words, and trust that, because it gives the right answer for the training texts, it will also give the right answer for the text of unknown authorship. The method I have used is simple – whether it is too simple is an open question – but it has the advantage that it is stable. [↑](#footnote-ref-3)
4. Matches between a play and itself are excluded from this test (they are not available anyway in my published data). [↑](#footnote-ref-4)
5. It is fair to acknowledge that some attributions are correct by only a fine margin. However, this problem seems to affect all attribution tests; or, at least, those for which the researchers choose to disclose the raw data that underlies their results. [↑](#footnote-ref-5)