# Summary

In part 1 we saw that formal N-grams other than 3-grams to 5-grams were not good for authorship attribution by the simple method I used. We also saw that no length of N-gram is good when using function word skip N-grams.

My purpose in part 2 is to go a little deeper into the data from part 1 and try to determine if all N-gram matches are better or worse than unique N-gram matches.

The analysis I have done, described below, suggests that unique formal N-grams are better, and we should use unique formal 3-grams and unique formal 4-grams in preference to the other kinds.

# Test Method

The raw data for 3-grams to 5-grams is here:[[1]](#footnote-1)



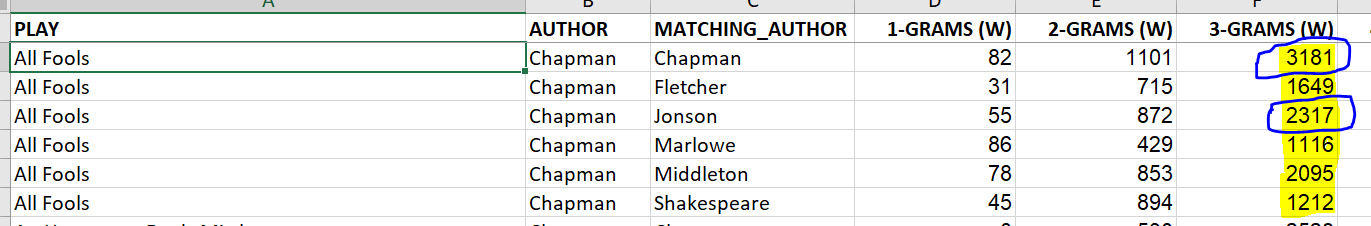
I have changed the spreadsheets of green/amber/red outcomes. The outcomes are still shown, and in the same colours, but each cell now displays a **winning or losing margin**, as a percentage. In a green cell, the margin is defined as:

This is the winning margin, telling us how comfortably the correct author won the attribution. The higher this margin is, the more confidence our method is expressing in the attribution. The highest possible winning margin is 100%, which will occur if the play has no matches at all with any other author.

In a red cell, the margin is defined as:

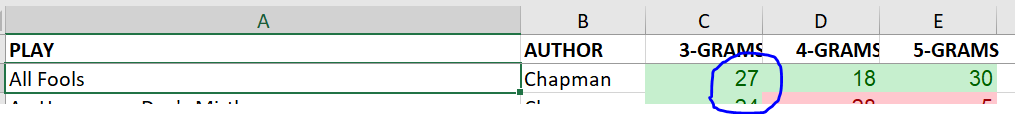
This is the losing margin of the correct author. The higher this value (in absolute terms, ignoring the minus sign), the more badly our method failed to attribute the play correctly (or our attribution was wrong). The worst possible losing margin is -100%. I round all percentages to the nearest integer.

For example, the raw data in part 1 for unique N-grams looks like this, with the 3-grams data for Chapman’s *All Fools* highlighted in yellow:



I have circled the highest and second-highest raw values in blue. Using the formula above, we can calculate Chapman’s winning margin as:

The new summary spreadsheet I provide in this part shows this percentage:



# Test Results and Interpretation

The new summary spreadsheets are given below, followed by my interpretation of the results.



Let us consider the all N-grams results first. Part 1 showed how badly 3-grams did compared to 4-grams: they got only 68 out of 86 attributions correct, whereas 4-grams got 82 out of 86 correct. Now that the losing margins are displayed, we can see that the failure was not as bad as it looked, as most of the losing margins are small; for example, *Monsieur D'Olive*, where Chapman lost by only 1%. Having said that, some of the winning margins are also very small, most notably for Jonson’s *The Devil Is An Ass*, which he won by such a small margin that it rounds to 0 and therefore shows up as amber on the spreadsheet. Nevertheless, the all 3-gram results still cannot be considered good.

Turning to the unique N-gram results, we can see that the 84 out of 86 attributions correct that the unique 3-grams got in part 1 was slightly flattering: *All’s Well That Ends Well* was attributed to Shakespeare by such a small margin that it rounds to 0 and now shows up as amber (the tie is with Jonson, not Middleton, as the raw data shows). Very surprisingly, unique 4-grams had failed to attribute *A Midsummer Night’s Dream* to Shakespeare. We can now see that he lost by only 6% (to Chapman).

In the unique N-gram results, observe, for the two *Tamburlaine* plays, how close the winning margins are to 100%. Observe also that although *The Faithful Shepherdess* is comfortably attributed to Fletcher, the attribution is less comfortable than that for his other plays, confirming the impression that it is the least Fletcherian of his plays.

In part 1, *The Blind Beggar of Alexandria* was the only play not to get any green cells in the unique N-gram results for any value of N. It received just one green cell in the all N-gram results, for 4-grams; but now that we have the winning margins displayed, we see that Chapman won it by only 2%. Someone should take a close look at that play.

The table below shows the average of the winning or losing margins for each kind of N-gram.

|  |  |  |  |
| --- | --- | --- | --- |
| Average margin (%) | 3-grams | 4-grams | 5-grams |
| All N-grams | 10 | 25 | 37 |
| Unique N-grams | 36 | 40 | 43 |

Unique 3-grams and unique 4-grams have average margins which are far ahead of those for all N-grams, while unique 5-grams are just a little ahead. We can therefore say with some confidence that unique N-grams are better than all N-grams for this method.

Unique 3-grams and unique 4-grams have higher success rates than unique 5-grams, both getting 83 out of 86 attributions correct, while unique 5-grams get only 75 out of 86 correct. This suggests that unique 3-grams and unique 4-grams are the best to use with this method.

# Why are Unique N-grams Better?

Why are unique N-grams better than all N-grams? Perhaps the answer is obvious. Authorship attribution is about working out what makes an author distinguishable. Elliott and Valenza use the terms ‘difference-mining’ and ‘resemblance-mining’. The essence of any attribution method is first to find out what makes an author different from his peers, to build up an author profile, and then see which other plays resemble that profile. What makes an author different is, by necessity, rare; and uniqueness is an extreme form of rarity. Perhaps that is explanation enough why unique N-grams provide the most accurate attributions.

There is the appearance of a paradox here. Our general understanding of scientific methods is that the greater the amount of data you have, the more reliable your conclusions will be. All N-grams provide a vastly greater amount of data than unique N-grams do, so why should they not be more reliable authorial markers? The answer is that, particularly for lower values of N, the number of formal N-gram matches is so super-abundant that common phrases, which everyone used, drown out the distinctive ones and confound our attempts to discover what makes an author different.

# A Darwinian View of this Method

We can now reflect on this simple method, introduced in part 1 and refined here. It is essentially a consistency checking method. You give it a complete set of attributions and it tells you how consistent they are with each other. Compare, for example, with Zeta, which is a binary choice method: you give it two alternatives, the base and the counter, and it picks the one that your unknown text is most similar to. This method avoids the problem inherent in any method that aims to attribute one text at a time, that the attribution may be a well-fitting one for that text but may not make good sense when introduced into a set of other attributions. This method insists on working with a set of attributions, checks each one against every other one, and asks if they make sense as a set.

It is even possible to imagine a process in which this method is seeded with attributions chosen arbitrarily and is then left to run by itself for a long time. The method would randomly vary attributions one at a time and, by a process akin to natural selection, eventually arrive at the best possible attributions. It would retain or discard mutations according to a rule that assesses their fitness. That rule might be a simple one, to just look at whether the mutation increased or decreased the number of green cells; or it might be a more sophisticated rule that takes winning and losing margins into account. An even more sophisticated form of the method might deal with the redness of an attribution by automatically disintegrating the play to see if the separate parts then become green. The process would finish when every cell is green, although that might only happen, if at all, after an unbelievable amount of disintegration.

# Maximal vs. Formal Matches

The conclusion I reached above was based on my published formal N-gram counts and is therefore valid only for them. The picture looks very different for maximal N-grams. Unique maximal N-gram matches are very few, as my N-gram summary files published in 2017 confirm. In theory they are very useful for authorship attribution; in practice, there are just not enough of them to provide a meaningful set of data. The results I have obtained above from unique formal N-grams should not be taken as implying that unique maximal N-grams are equally useful, if at all.

Pervez Rizvi

October 2018

1. In part 1, I used only the division of *King Lear* that excludes the Folio-only lines. This time, I have used the whole play. This means that a few of the numbers differ by small amounts, but not by enough to swing the attributions. [↑](#footnote-ref-1)