# Whole Canons Method

This document shows you how to use the ‘whole canons method’[[1]](#footnote-1), which is a way of testing authorship attributions for their consistency with the evidence of the formal N-gram matches between plays.

The method was developed in my two-part article “Which N-grams are the Best?”. I have not repeated the definition here.

# Why use it?

When you make an authorship attribution, you enlarge the canon of the author to whom you attribute the play and diminish the canon of the author to whom it was formerly attributed. Either or both of those changes could affect the safety of a different attribution. If a play contains a set of words or phrases in large quantities, perhaps because of its subject matter, then moving it from one author to another may materially change our counts of how often each of those authors uses those words or phrases. It is conceivable that some existing attribution, perhaps a marginal one, might depend on those very counts, and might thereby become either more, or less, safe. Unless you check, you cannot be sure. The checking is not always done. I have formerly pointed out the same problem with compositor attributions.

This method has the advantage that it tests all attributions equally and at once. It can therefore be used at any time to see how compatible new attributions are with existing ones. The information it gives may sometimes draw attention towards attributions that are marginal and deserve closer study, or it may even suggest new attributions.

# Installation

Open the ZIP file provided and extract all the files into a folder of your choice. You may extract different files into different folders if you wish.

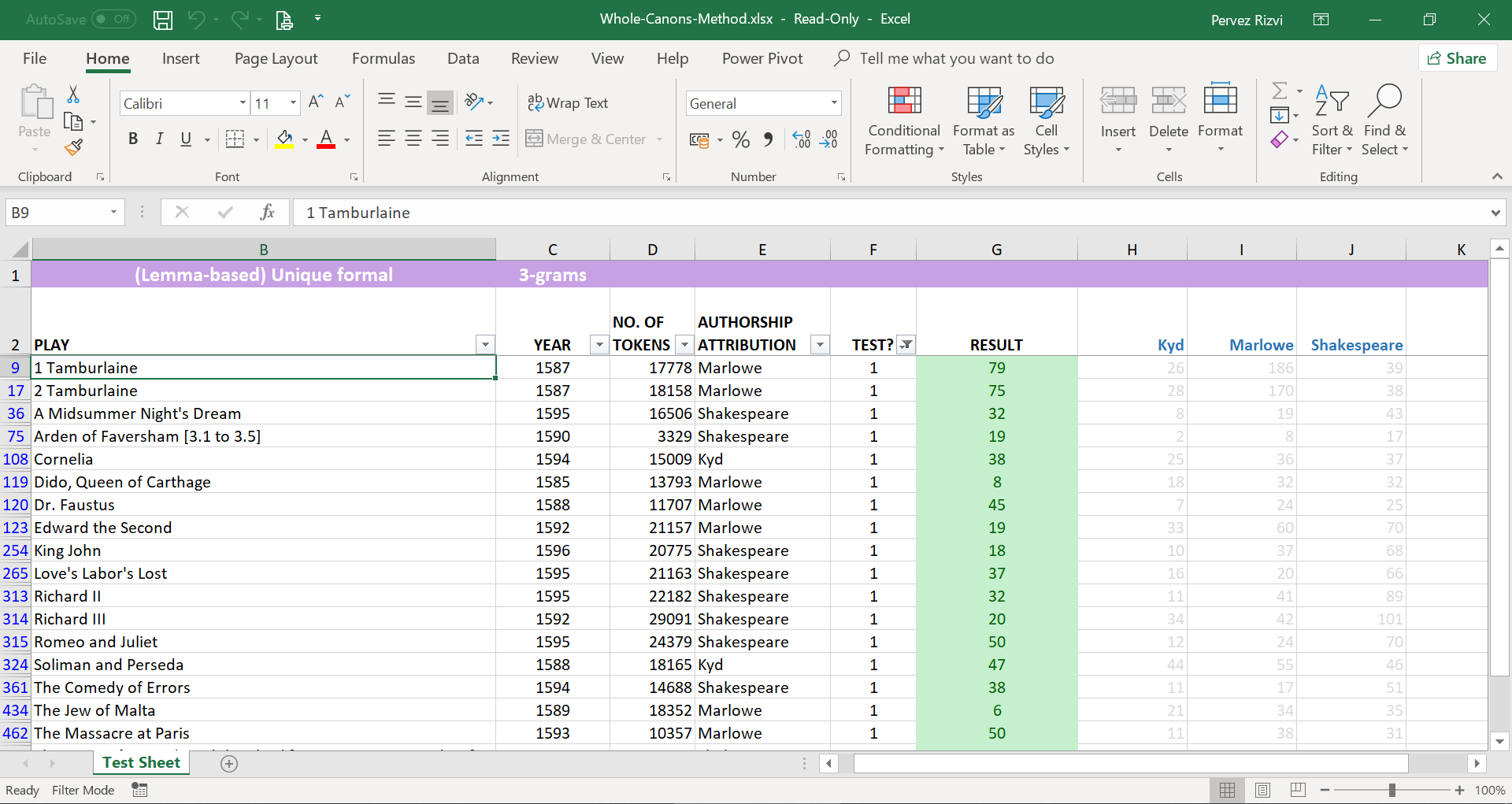
Among the files provided, you will work with the one called **Whole-Canons-Method.xlsx**. I shall refer to this file hereafter as ‘the spreadsheet’.

You do not need to look in the other files, which I shall refer to as ‘data files’. They contain the raw counts that the tests use.

# Quick Start

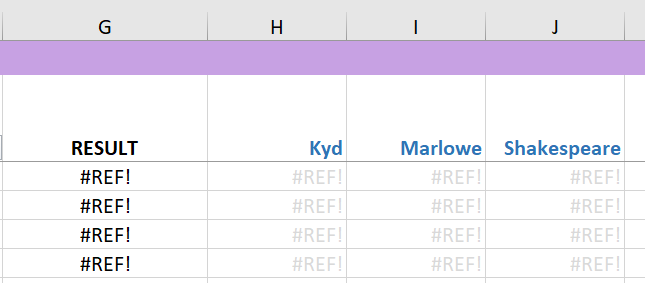
This section is intended to allow you to get started with the method quickly, using a ready-made attribution test, to get a good idea of what it does.[[2]](#footnote-2) You may then follow the more detailed instructions below to do your own choice of tests.

Close all other spreadsheets that you have open. Open the spreadsheet. You should see a screen like this[[3]](#footnote-3):

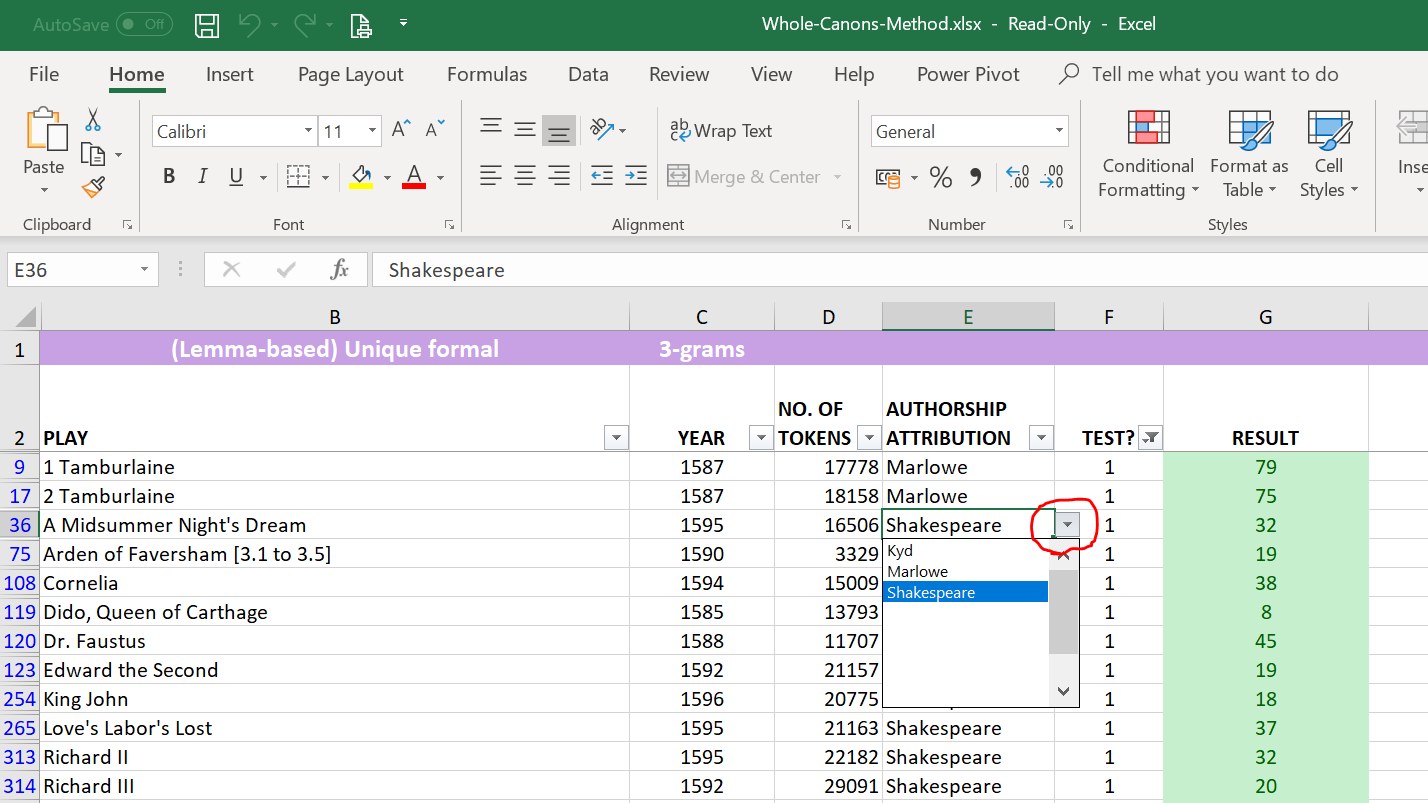


This screen is showing you the results of a test already done on a small set of plays. The top line, in lavender colour, tells you that you are using lemma-based counts of unique formal 3-grams.

If, on the other hand, you see something like this, then don’t worry – just carry on with the instructions and it will be fine:

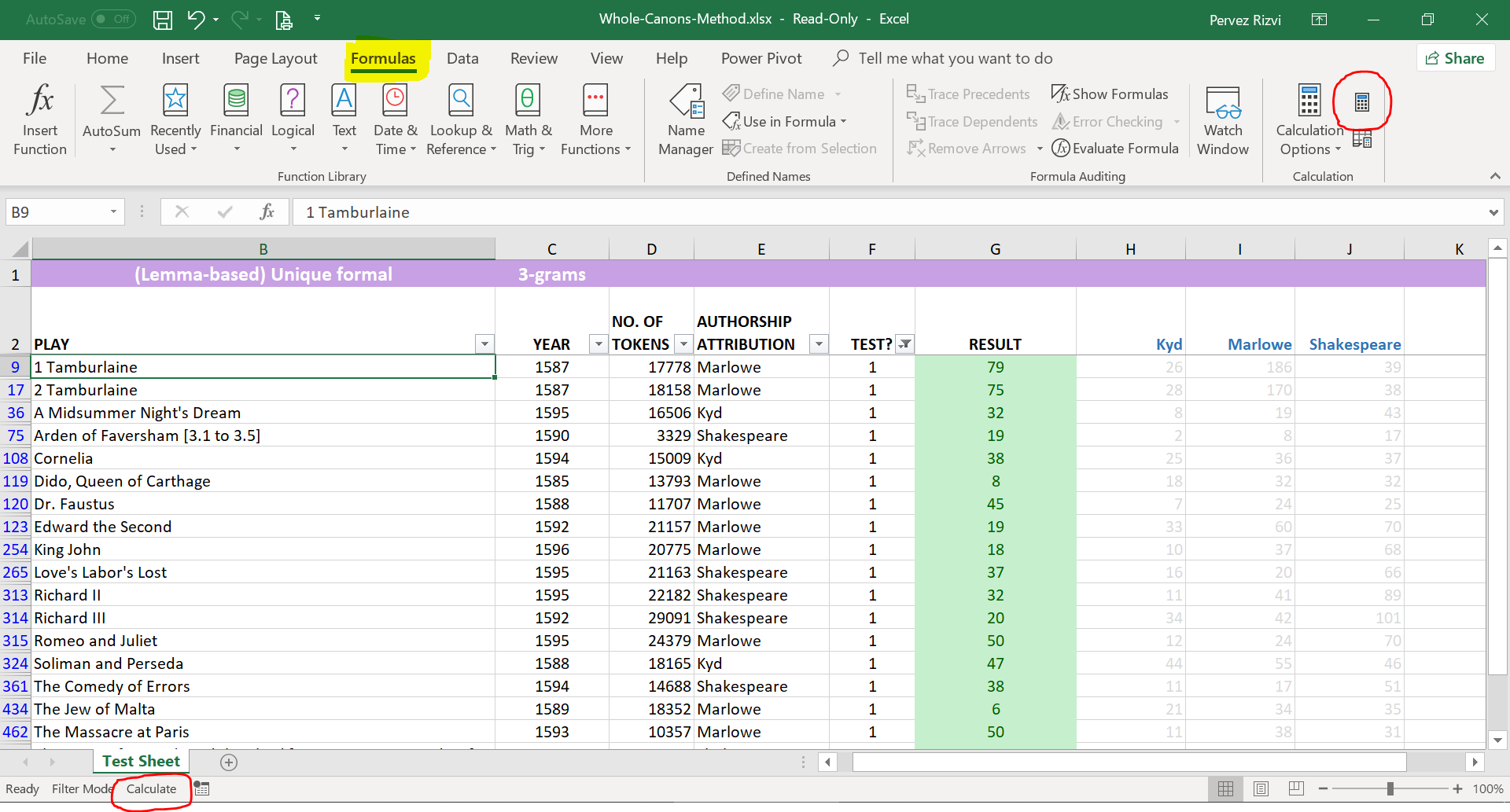


Use the “Authorship Attribution” column to change some of the attributions, using the dropdown lists provided, as shown below:

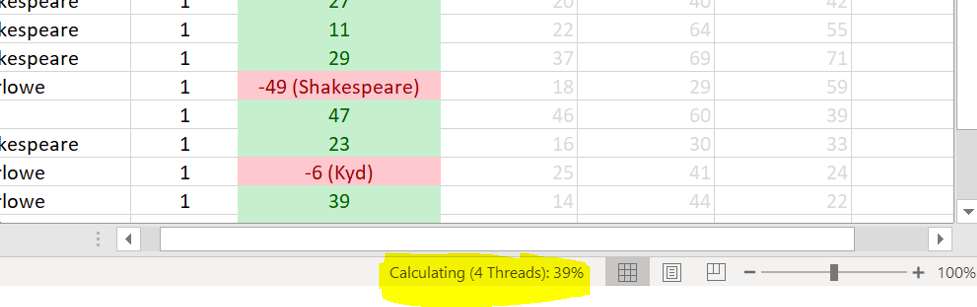


When you have finished making your attributions, or before then, open the spreadsheet called **Unique-L.xlsx**. Having opened it, you can minimize it, but do not close it.

Calculate the test results by pressing either the Calculate button on the top-right or the bottom-left, both circled in red below:



Wait a few seconds for Excel to do the calculations. While they are being done, you will see a progress bar at the bottom right of the window:



When the calculations are complete, the progress bar will disappear, and you will see the results of the test.

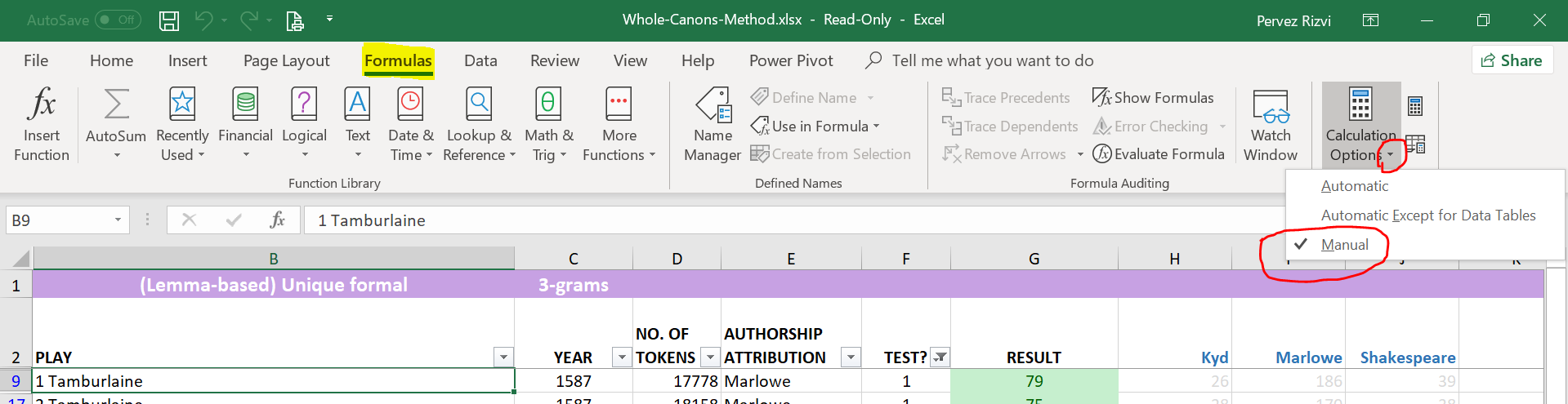
# How to do your own Tests

**General Warning**. Do not copy a cell and paste it into another cell; instead, type what you want into each cell, or select from the dropdown list, or delete its contents using the delete key. If you copy and paste, you may corrupt the formulae in the spreadsheet because there are so many cross-references between cells.[[4]](#footnote-4)

Close all other spreadsheets that you have open. Open the spreadsheet and follow the instructions below.

## Turn off Auto-Calculation

On the Formulas menu, click the Calculation Options arrow as shown:



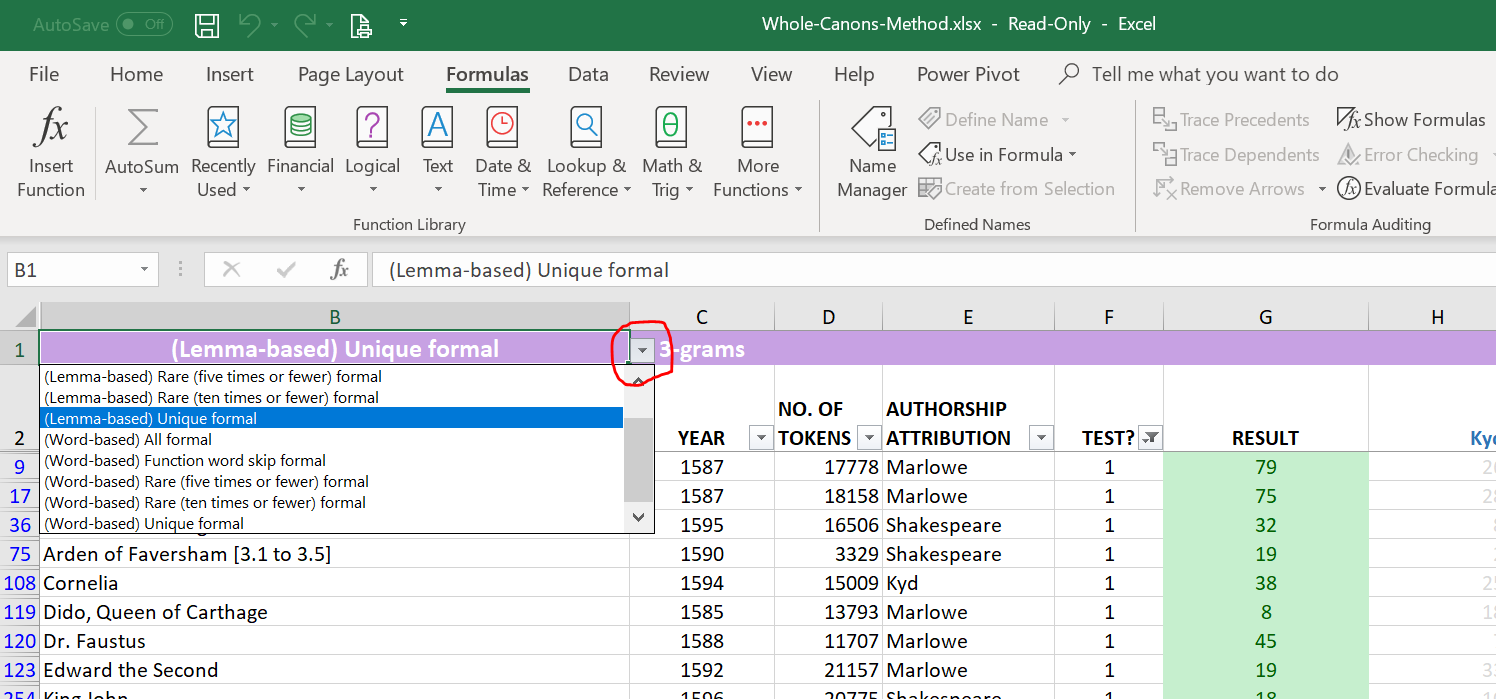
Check that the Manual option is ticked as shown.

It is not essential to do this. However, if you do not do it, then Excel will perform recalculations every time you change a cell, which is likely to slow you down quite a bit.

Be aware that if you have other spreadsheets open that have automatic calculation enabled – which is the default in Excel – then this spreadsheet will also perform auto-calculation when you open it. That is why it is good practice to close all other spreadsheets before you start these tests.

## Choose the Counts

Click in the cell on the top-left of the sheet, and then click on the arrow shown in red and choose the kind of counts you want to use in your tests.



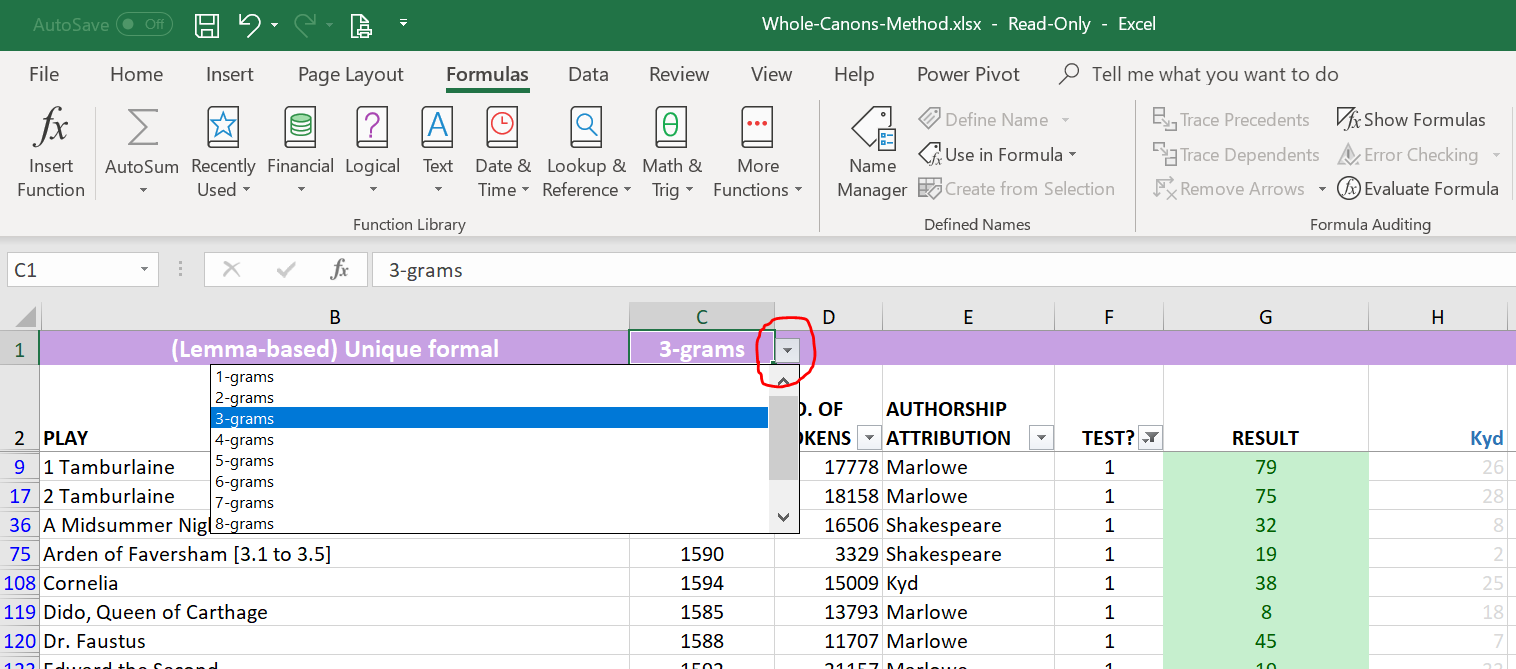
You have a choice of ten kinds of N-gram match counts, five of which are calculated using lemma-based matches, and the other five using word-based matches.[[5]](#footnote-5)

A ‘unique’ match is defined as one that occurs in exactly two *plays*.[[6]](#footnote-6) A ‘rare (five times or fewer)’ match is defined as one that occurs in five or fewer *plays*.

Having made your choice, open the corresponding data file, the names of which are self-explanatory. For example, if you have chosen to use lemma-based unique formal counts, then you must open the **Unique-L.xlsx** file, leave it open and minimize it. If you forget to do this, the calculations will give errors later.

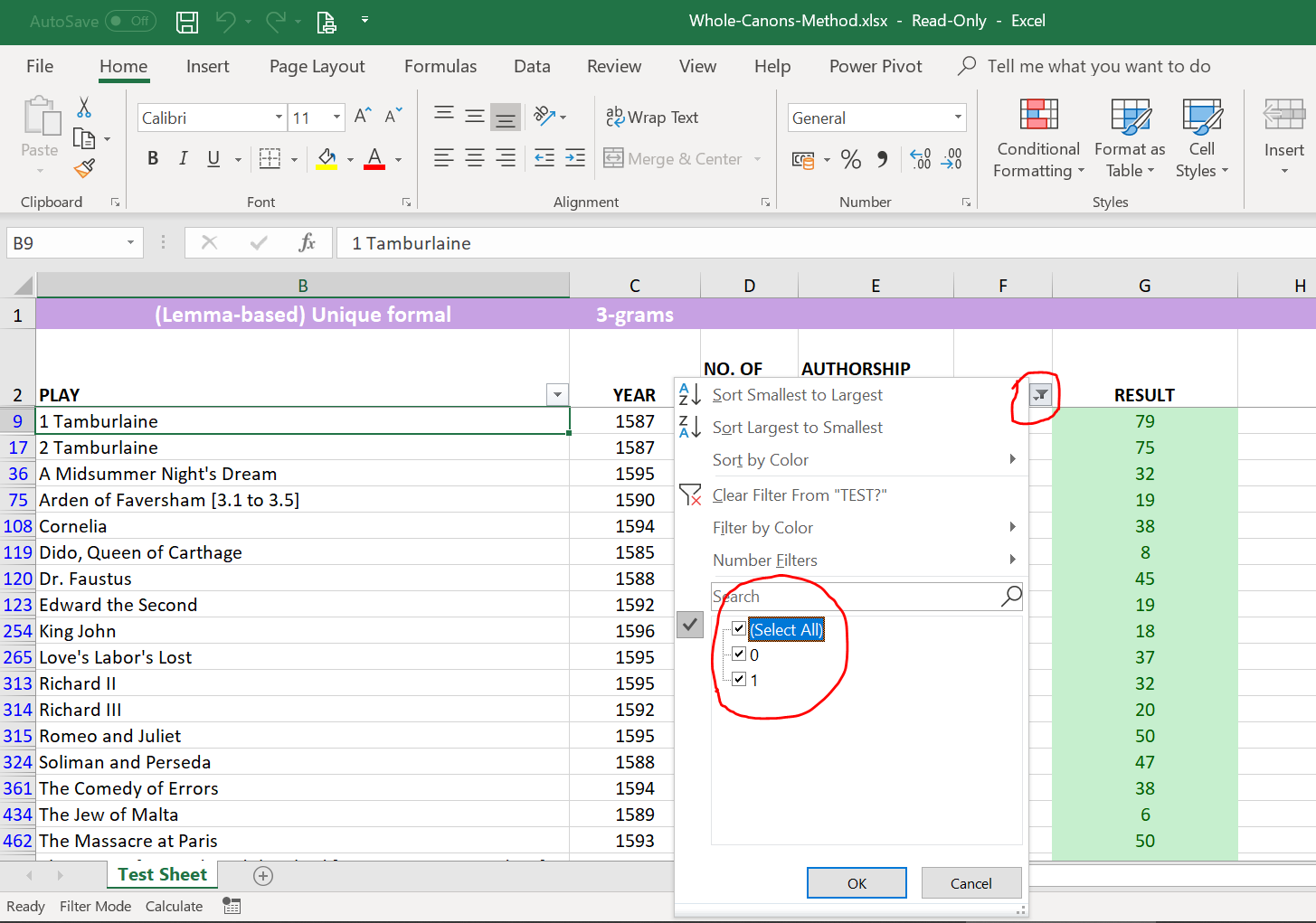
## Choose the N-grams

In the cell to the immediate right, click on the arrow circled in red and choose the length of N-gram that you want to use in your tests:

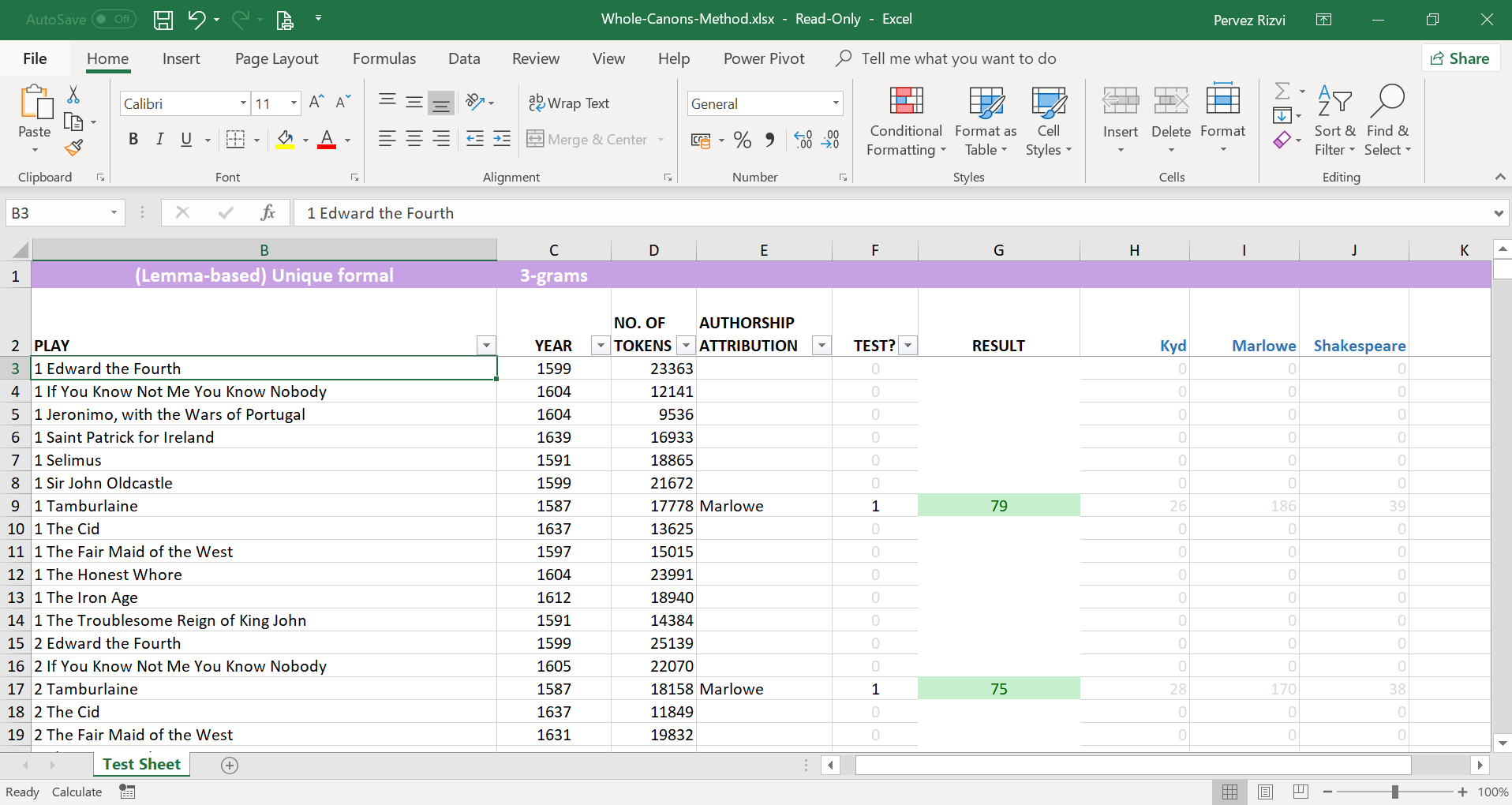


## Choose your Plays

Change the filter in the “Test?” column to show all plays:



You should now see all plays:



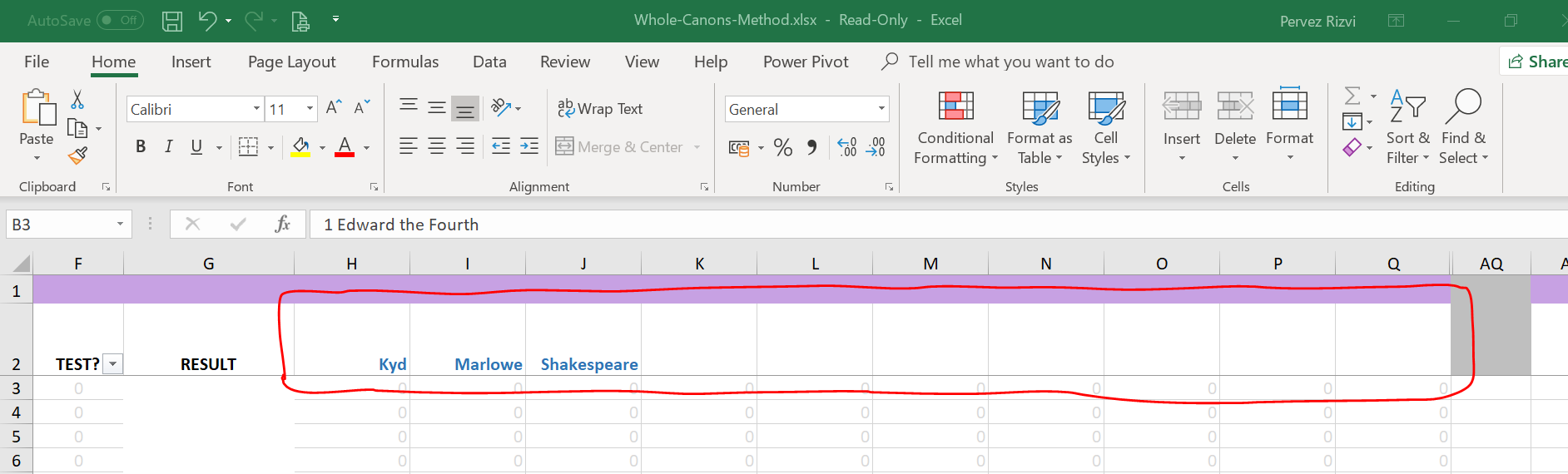
Find the plays you want to test and enter 1 in the “Test?” column for each of them. You can search for the play names or use the filters in the “Play” or “Year” columns.

If some plays already have a 1 against them in that column but you do not want to include them in the test, then change the 1 to 0.

When you have finished choosing the plays, filter the “Test?” column again to show only the plays with 1. This will allow you to check that you have not inadvertently included some plays that you did not intend to include.

## Enter your Authors

Look at cells H1 to Q1, as shown below:



This area is where you enter, in a horizontal line, the names of the authors to whom you will attribute plays. You are limited to ten authors, but this limitation can be removed once you become an advanced user of the spreadsheet.

It is not essential to enter ten authors; you may enter fewer. You can enter them into any of the ten cells and in any order. Do not enter the same author twice, otherwise you will get an error at the calculation stage.

## Make your Attributions

Do this as explained above in the Quick Start section above. You do not have to assign a play to every author you entered above.[[7]](#footnote-7)

## Get the Results

Get Excel to perform the calculation, as explained in the Quick Start section above.

Having got the results, you can experiment with them.[[8]](#footnote-8) For example, you could change the length of the N-grams you are using; or change some attributions, perhaps by entering new authors. You could even change the kind of counts you are using; for example, you could switch to using all N-grams instead of unique N-grams. If you do that, then make sure you open the corresponding data file before you attempt recalculation.

When you obtain a result you want to keep, you can save the spreadsheet under a name that describes the test you did. Or you can print the sheet, which is formatted to print on A4 paper.

# Advanced Use

## Editing the Spreadsheet

The spreadsheet is protected, to prevent you from corrupting it by accidentally overwriting one of the formulae. If you are confident using Excel and want to change the spreadsheet – for example, to change the weighting formula – then you can unprotect the sheet, for which you need to know the password. The password for the spreadsheets and the data files is **Shakespeare**

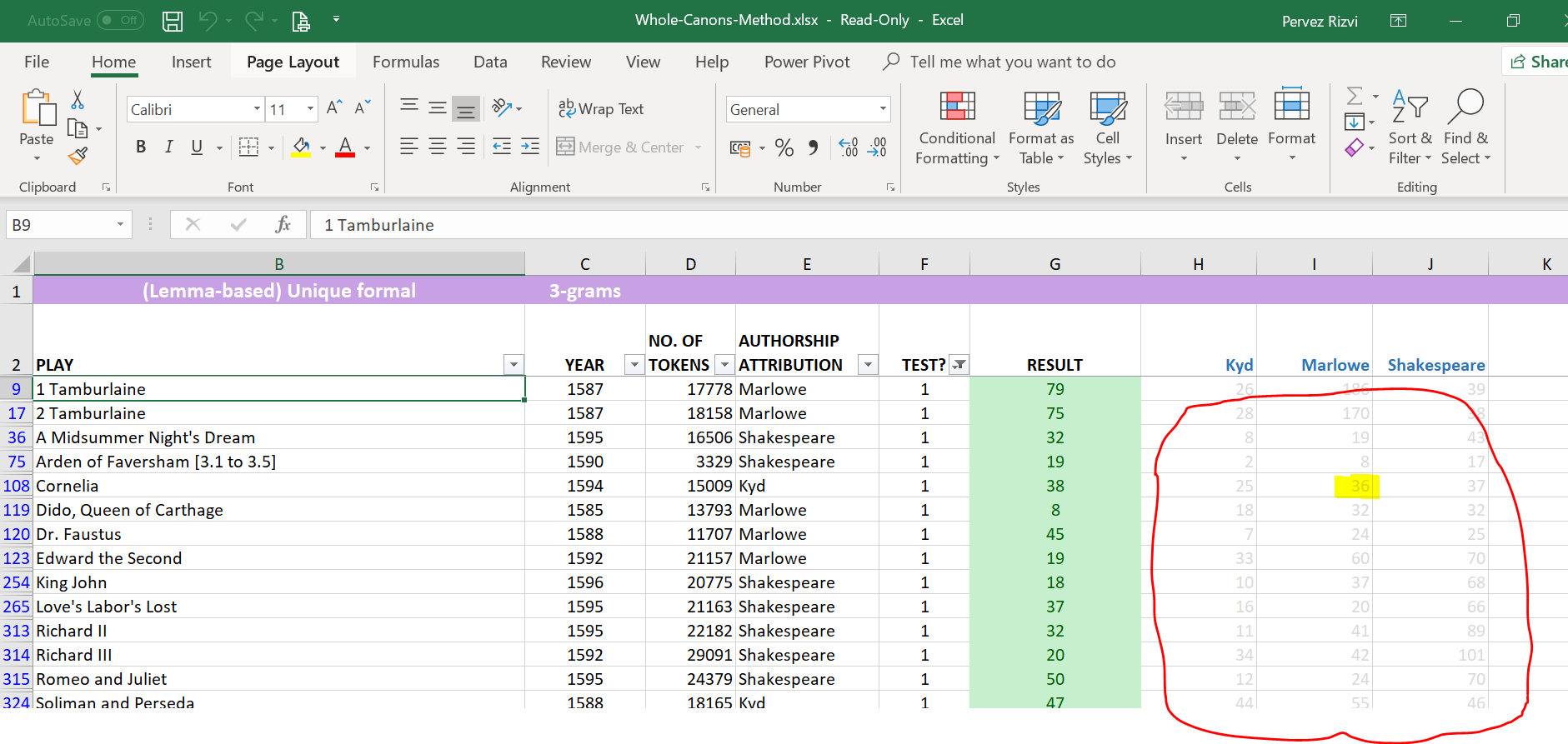
## Tests for New Divisions of Plays

You can only do tests on the divisions of plays for which I have provided the counts. For example, if you believe that *The Massacre at Paris* is co-authored and want to test that theory, then you cannot do it, because I have only provided counts for the whole play, not divisions of it.

If you want to do this, you will need to contact me to tell me how you want to divide the play, so that I can produce separate counts for each division and make them available to you.

## Looking at the Raw Data

When you have calculated the results of a test, you will see numbers like the ones circled below, under the names of your authors:



These are the totals of the number of matches between the play on this row and the author in this column. For example, the number 36 highlighted in yellow is the number of lemma-based unique formal 3-gram matches (unweighted) between *Cornelia* and the Marlowe plays in the test.

It is worth keeping an eye on these numbers. An unusually large count might indicate a strong association between a play and an author, but it might also indicate that one play by that author happens to have lots of occurrences of the same matching phrase (you could test that by removing some plays from that author’s canon and seeing how the count goes down). If all the counts you see are very small, it will be a sign that there is not enough data for a reliable test.

These counts are greyed out in the spreadsheet. If you wish, you can make them easier to read by making them black. To do that, you will first need to unprotect the spreadsheet, taking care not to change any formulae. Protect the spreadsheet again when you have finished.

## Making Deletions

If there are plays that you think you will never want to use in your tests then you can simplify the spreadsheet, and make calculations faster, by deleting their rows. Do not delete any single cells; always delete the whole row, which is safe to do.

Similarly, if there are N-grams, such as 10-grams, which you think you will never want to use, then you can save some disk space by deleting their columns from the data files.

## Adding Columns

If you are confident that you have understood how the formulae work, then you can add extra columns of information from your own sources, so you can take them into account in the test, or you can add columns to allow more than ten authors to be used in a test. You may need to adjust some formulae if you do this, so make sure you have understood them first.

# Are these Counts Actually Correct?

Yes.

The results you get are only as reliable as the counts I have provided. Three of the ten data files contain the counts I published in June 2018. At that time, I provided Concordance files, with instructions on how to use them to check the counts, to trace them back to the matching phrases in the original texts.

For the seven kinds of counts provided for the first time today, I have not yet produced the Concordance files, but I shall produce them in a few weeks and make them available. The new counts are produced by the same program that produced the old ones and, in advance of the Concordance files, I have manually checked some counts from each data file against my database of plays and found them to be correct.

Pervez Rizvi

2 December 2018

1. The method was referred to by this name in passing by MacDonald P. Jackson, who was kind enough to comment on it. I have gratefully adopted the name. [↑](#footnote-ref-1)
2. The ready-made test included as a sample in this spreadsheet was suggested by MacDonald P. Jackson. Its idea is to test together the early sole-authored Shakespeare plays, the Marlowe plays, the three plays generally accepted as being by Kyd, and the Countess scenes in *Edward III* and scenes 4-8 in *Arden of Faversham*. [↑](#footnote-ref-2)
3. My screenshots are from Excel365. The spreadsheet should work in any recent version of Excel, though the screens might look different for you. [↑](#footnote-ref-3)
4. There seems to be no way to prevent this without introducing a macro into the spreadsheet, which I am reluctant to do as it causes Excel to issue warnings that may dissuade people from using it. [↑](#footnote-ref-4)
5. For example, ‘glassy stream’ in *Hamlet* will match ‘glassy streams’ in *1 Henry VI* when using lemma-based matching, but not when using word-based matching. [↑](#footnote-ref-5)
6. If a play is divided into two or more divisions, and the matching phrase is found in more than one of those divisions, then the phrase is still treated as a unique match, provided it is found in exactly one *other* play. This definition of uniqueness ensures that the counts remain constant, even when different scholars divide the same play into different divisions. [↑](#footnote-ref-6)
7. Having made your attributions, you could of course go back and change some of the author names you entered earlier. For example, you could attribute a play to Kyd and then delete his name from where you entered it earlier. If you do this, you will get an error at the calculation stage. [↑](#footnote-ref-7)
8. If you wish, you can now replicate the tests I did in my series of articles starting with “Which N-grams are the Best?”. You will notice some minor differences in the percentages. These are for two reasons. In the first article, I used only the division of *King Lear* that excludes Folio-only lines; in subsequent articles I used the whole play. The data files provided now contain counts for the whole play. Secondly, in the results published earlier, I had rounded at two points in the calculation. The spreadsheet presented in this article rounds only at the end of the calculation, so its results are a little more accurate. [↑](#footnote-ref-8)