## Introduction

All matches in this section of the downloads page are Formal matches, as defined [here](http://shakespearestext.com/can/counting.htm).

I have made counts by matching tokens and types separately. This means that if a phrase occurs 3 times in one play and 2 times in another, then it contributes 6 to the count of matches by token between those plays but contributes 1 to the count of matches by type. Type-based matches are provided for completeness only; I am not aware of their being used by any scholar.

All N-gram Matches are what are traditionally understood as N-gram matches. They are matches of all N-grams, looking at **all** words in all texts, except speech prefixes, for values of N from 1 to 10.

Unique N-gram Matches are a subset of all N-gram matches, containing only the matches that occur in exactly two plays. If an N-gram occurs in exactly two plays, then that does not prevent it from being counted as unique even if it occurs several times in those plays. There is a further stipulation. Suppose we have divided a play into several divisions. If an N-gram occurs in one of those divisions and in just one other play, then it is counted as a unique match even if it also happens to occur in some other division of the same play. This stipulation is made to ensure that this list remains constant, no matter how or whether we choose to divide a play into divisions. More technically, uniqueness, as defined here, is *invariant under division*.

## File Structure

|  |  |  |
| --- | --- | --- |
| Excel Column | Name | Meaning |
| A | Play | The play or division whose search results are summarised in this file. This is of course the same in every row. It is provided to make it easier to merge rows from different files without losing the knowledge of which file they came from. |
| B | Year | The year, according to the SHC and Folger websites, in which the play was written (which may be much earlier than the date of publication). These dates are not always reliable. The dates in the latest authority (Wiggins) may differ. |
| C | Tokens/Types | The number of tokens or types in the play or division, depending on whether this is a file of token counts or type counts, excluding speech prefixes but including stage directions. |
| D | Matching Play | The play or division whose matches are summarised in this row. |
| E | Matching Play Year | The year, according to the SHC and Folger websites, in which the matching play was written (which may be much earlier than the date of publication). These dates are not always reliable. The dates in the latest authority (Wiggins) may differ. |
| F | Tokens/Types in Matching Play | The number of tokens or types in the matching play or division, excluding speech prefixes but including stage directions. |
| G to P | N-grams | The counts of N-gram matches between the two plays or divisions listed on this line, for values of N from 1 to 10. |
| Q to Z | N-grams Weighted | The values in columns G to P, weighted by being divided by the sum of the number of tokens or types in the two plays or divisions. For example, Q = G/(C+F). |

The rows in the Summary file are sorted by column T, the weighted number of 4-gram matches. You are free to sort the rows in a different order. If you do, then remember that you must tell Excel to sort in descending order (“Largest to Smallest”), as by default it sorts in ascending order.