# Summary

In this article I do a control test on the method I have used in the articles “Which N-grams are the Best?” and find the method to be unbiased.

# Test Method

The idea is that we attribute plays to authors at random and then see if the method accepts or rejects the attributions.

For our standard 86 plays we have six available authors. Therefore, if we assign plays to authors at random, we should expect to get about one-sixth of our attributions correct. If our method is unbiased, it should accept approximately that proportion of our random attributions and reject the rest.

To avoid suspicions that I have fixed the results, I chose a pseudo-random rather than a truly random method of attribution. I ordered the 86 plays in order of the number of words they contain (see the Appendix for this order). I then went through the list in order, assigning plays in turn to each of the six authors, in alphabetical order of their names. In other words, Chapman gets the first play on the list, Fletcher the second, Jonson the third, Marlowe the fourth, Middleton the fifth, Shakespeare the sixth, and then Chapman gets the seventh, and so on. Since 86 is not an exact multiple of 6, it means that four authors get 14 plays each and the other two get 15 each.

I have tested unique formal 3-grams and unique formal 4-grams, as those are the ones found to be the best in my earlier articles.

# Test Results

I wrote “about one-sixth” above as the number of attributions we expect to be accepted by the method in this test, but we can calculate the number more exactly, using the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| Author | Actual No. of Plays by Author (A) | No. of Plays ‘Randomly’ Assigned to Author (B) | Expected No. of Accepted Attributions () |
| Chapman | 13 | 15 | 2.27 |
| Fletcher | 14 | 15 | 2.44 |
| Jonson | 13 | 14 | 2.12 |
| Marlowe | 6 | 14 | 0.98 |
| Middleton | 12 | 14 | 1.95 |
| Shakespeare | 28 | 14 | 4.56 |
| Total | **86** | **86** | **14.32** |



We see that unique 3-grams get 14 out of 86 attributions correct, while unique 4-grams get 11 out of 86 correct. These numbers are close to the 14.32 that we should expect, allowing for the usual variation we get in real-life data of this size, compared to the theoretical values. The method appears not to be biased. This means that when we get high success rates, such as 83 out of 86 correct, it means that our attributions have a genuine value and are not simply what the method was pre-disposed to give us.

# Appendix – No. of Words per Play

These counts include stage directions but exclude speech prefixes:

|  |  |
| --- | --- |
| No. of Words | Play |
| 10357 | The Massacre at Paris |
| 11707 | Dr. Faustus |
| 13278 | The Blind Beggar of Alexandria |
| 14688 | The Comedy of Errors |
| 14743 | The Alchemist |
| 16452 | The Conspiracy of Charles Duke of Byron |
| 16506 | A Midsummer Night's Dream |
| 16545 | An Humorous Day's Mirth |
| 16555 | The Chances |
| 16630 | The Tempest |
| 17077 | The Two Gentlemen of Verona |
| 17267 | Caesar and Pompey (The Wars of Caesar and Pompey) |
| 17274 | A Chaste Maid in Cheapside |
| 17559 | Monsieur D'Olive |
| 17778 | 1 Tamburlaine |
| 17885 | A Game at Chess |
| 18158 | 2 Tamburlaine |
| 18271 | A Mad World, My Masters |
| 18352 | The Jew of Malta |
| 18431 | A Trick to Catch the Old One |
| 18756 | The Tragedy of Charles Duke of Byron |
| 18759 | Rule a Wife and Have a Wife |
| 18903 | The Mad Lover |
| 18905 | Catiline His Conspiracy |
| 18910 | The Puritan, or The Widow of Watling Street |
| 18950 | All Fools |
| 19126 | The Widow |
| 19198 | More Dissemblers Besides Women |
| 19579 | Julius Caesar |
| 19827 | Michaelmas Term |
| 19851 | Twelfth Night |
| 19906 | Bussy D'Ambois |
| 19971 | The Phoenix |
| 20004 | The Revenge of Bussy D'Ambois |
| 20161 | The Revenger's Tragedy |
| 20359 | Your Five Gallants |
| 20361 | The Faithful Shepherdess |
| 20651 | Sir Giles Goosecap |
| 20753 | Women Pleased |
| 20775 | King John |
| 20778 | Bonduca |
| 20929 | The Taming of the Shrew |
| 21113 | The Gentleman Usher |
| 21124 | Much Ado About Nothing |
| 21157 | Edward the Second |
| 21163 | Love's Labor's Lost |
| 21248 | The Pilgrim |
| 21267 | The Merchant of Venice |
| 21536 | The Merry Wives of Windsor |
| 21572 | The Widow's Tears |
| 21604 | The Wild Goose Chase |
| 21730 | As You Like It |
| 21869 | A Wife for a Month |
| 22086 | The New Inn, or The Light Heart |
| 22182 | Richard II |
| 22717 | The Island Princess |
| 23026 | All's Well That Ends Well |
| 23191 | May Day |
| 23243 | The Woman's Prize, or The Tamer Tamed |
| 24379 | Romeo and Juliet |
| 24642 | The Humorous Lieutenant |
| 24642 | Henry IV, Part 1 |
| 24682 | Antony and Cleopatra |
| 24937 | Every Man in His Humour |
| 24963 | The Winter's Tale |
| 24970 | Valentinian |
| 25439 | The Staple of News |
| 25642 | Sejanus His Fall |
| 25845 | King Lear |
| 25983 | Troilus and Cressida |
| 26123 | Epicoene or The Silent Woman |
| 26126 | The Loyal Subject |
| 26146 | Henry V |
| 26210 | Henry IV, Part 2 |
| 26304 | Othello |
| 26821 | Volpone |
| 27287 | The Devil Is an Ass |
| 27513 | Cymbeline |
| 27656 | Coriolanus |
| 28787 | Poetaster |
| 29091 | Richard III |
| 29451 | Women Beware Women |
| 30305 | Hamlet |
| 32749 | Cynthia's Revels |
| 36388 | Bartholomew Fair |
| 37978 | Every Man Out of His Humour |