PART THREE

New directions

CHAPTER 3.1

Shakespeare and authorship attribution methodologies

HUGH CRAIG

This chapter is a primer in Shakespeare authorship attribution.[1] I present a series of examples to illustrate some of the key considerations which one should bear in mind in an attribution study. They have some technical aspects which will require patience on the reader's part to work through but doing so will (I hope) help to acclimatize new and prospective attributionists to the constraints and opportunities of this practice.

On the whole the questions I deal with are statistical rather than literary. We come to Shakespeare attribution because of an engagement with the content of his work and his contemporaries', but the skills and mental habits we need for quantitative attribution are not literary, or at least not part of the usual literary training. Those most interested in Shakespeare tend to focus on resonant details and linger on individual instances, and seek for large intuitive insights, but for attribution the key is even, wide attention and a systematic method inoculated against bias. The quantitative part of attribution occurs at the juncture between statistics and language. It fits awkwardly with the documentary and historical side of determining authorship, and with evaluative, interpretive literary studies. A quantitative approach is, unavoidable, nevertheless, if we are to make reliable assignations of anonymous and disputed texts and establish levels of confidence about an attribution.

My first two case studies concern the author effect on which attribution depends. I show that this effect is objectively present in Shakespearean works, and go on to discuss exceptions and limitations to consistent authorial self-resemblance. I then treat two other key concepts: the law of large numbers and the problem of overfitting. Finally, I discuss the practicalities of assembling a corpus and applying statistical procedures.

Some scholars are sceptical about the idea that playwrights of Shakespeare's era have distinctive styles which make their works recognizable as belonging to them and no one else. For some, the doubts are based on a belief that individual authorship was subservient to broader cultural forces in the early modern period (McMullan 2000: 6, 174, 193–5). For others, the problem is that variation within

authorial canons is too great (Rudman 2016). Near the beginning of his Oxford History of English Literature volume, *English Drama 1586–1642*, G. K. Hunter strikes a general cautionary note on the topic:

> From a seat in the stalls it is difficult to know why we should think [Ben Jonson's play] *The New Inn* is by the same author as *The Alchemist*, why the Heywood who wrote *A Woman Killed with Kindness* is the same Heywood who wrote *The Golden Age*, why the author of *The Merry Wives of Windsor* is also the author of *The Tempest*.
>
> (1997: 3n.5)

Authorship attribution has to start from a ground zero, which is the suspicion that in this period an author's plays have only weak connections to each other. If we had no external guide, if this was a blind tasting, as it were, would we connect two unlabelled plays by the same author with each other?

I first reframe the problem as the question of whether authors repeat characteristic distinctive phrases. Do they keep the same writing habits as time goes on, when they turn to different genres or try different topics? If not, there will be no basis to identify a mystery work as theirs. It will have no necessary connection to the works we know to be written by them. If so, we can go forward with some confidence that there is a degree of predictability about what an author writes, though this will not remove the constant problem of defining just how much predictability, and in what circumstances.

I also shift the ground from an audience's perceptions of plays performed to patterns in written texts detectable by a computer programme, and test the conundrum of authorial self-consistency versus authorial indeterminacy on a numerical basis. Are an author's works more like each other than they are like other authors' works, even if those other works share a genre, or an era in time, or a plot outline? We have to be careful that any test is as fair as possible. Either side of the argument has to have an equal chance to prevail. I will now present an attempt to do this.

We take five playwrights of the time, including Shakespeare, and works by them with a date of first performance between 1580 and 1619, comfortably straddling Shakespeare's career. For the dating of the plays we rely on Wiggins and Richardson's *Catalogue of British Drama* (2012–). We make a random selection of nine plays from each, and then choose a further two plays from each writer as test plays, again using a random selection if there are more than two to choose from.[2]

We focus on repeated sequences of six words. Choosing a sequence length of six is an arbitrary choice, designed to yield repetitions which will be unusual enough to give us only small numbers to deal with, so that each instance can be examined without undue labour. As a matter of simple mathematics, as the sequence length gets longer, there are fewer repetitions. Working in single words yields relatively few different words, most of them frequently recurring, two-word sequences have more different sequences, and fewer repeats, and so on.

We seek examples that are rare enough to be potentially characteristic of an author rather than belonging to the common parlance of the day. To determine

rarity, we require from the beginning that any sequence we include should not appear in either of two reference sets. One set consists of all the plays in the corpus which are dated 1580–1619 and attributed to single authors (other than our chosen five) in the *Catalogue*. The other is the full set of 1580–1619 plays by the four chosen authors other than the author set being used at the time to test authorial linkages. If a sequence satisfies the condition that it does not occur in the general single-author set, or in the appropriate four-author set, then we can be sure that it is tolerably rare, and we can attach some importance to any repetition.

We have ten test plays, two from each of the five authors. We find how many rare six-word sequences are shared between a given test play and each of the five nine-play authorial sets. Each play thus has five scores, one for each authorial set. There may be no matching rare sequences, or one, or more. The results are shown in Table 3.1.1, with test plays as rows and authorial sets as columns.

In the table, the cells showing the number of sequences shared by the test plays and the nine-play sets of their known authors are shaded. This number is always more than zero apart from one case, Chapman's *Conspiracy of Charles Duke of Byron*, which does not share a rare sequence with the Chapman set. The unshaded cells show the number of sequences shared between the test plays and the authorial groups other than that of their known author. These are zero in every case.

Table 3.1.1 Rare Six-word Sequences Shared Between Ten Plays and Five Authorial Sets

Test plays		Sets of nine plays as sources for matched 6-grams				
		Chapman	Fletcher	Jonson	Middleton	Shakespeare
Chapman	*Conspiracy of Charles Duke of Byron*	0	0	0	0	0
Chapman	*Revenge of Bussy D'Ambois*	6	0	0	0	0
Fletcher	*Monsieur Thomas*	0	7	0	0	0
Fletcher	*Mad Lover*	0	6	0	0	0
Jonson	*Bartholomew Fair*	0	0	16	0	0
Jonson	*Poetaster*	0	0	14	0	0
Middleton	*No Wit, No Help Like A Woman's*	0	0	0	23	0
Middleton	*Puritan*	0	0	0	15	0
Shakespeare	*Henry IV, Part Two*	0	0	0	0	7
Shakespeare	*Julius Caesar*	0	0	0	0	3

Authors do repeat themselves, according to this test. When we seek out parallels between the test plays and their known authors, we find (nine times out of ten) that there are links, sometimes many of them. In the extreme case of Middleton's *No Wit, No Help Like a Woman's*, there are twenty-three. When, under exactly comparable circumstances, we seek out parallels with other authors, we find none.[3]

We might have anticipated that any shared rare sequences would be evenly distributed across same-author and other-author sets, because of the ebb and flow of dialogue in a theatre with common preoccupations and ways of making drama, but this is not the case. We might have thought that there would be no such shared sequences, because any new sequences in a given test play would be unique, given the wide range of possibilities available in a language like English, and because of the motivation of authors to provide something new for audiences. This is not true either. It turns out that authors have a tendency to resort to characteristic unusual phrasings even in a much later play, or in a play in a different genre.

Early in *Julius Caesar* Caska says to Cassius, 'Stand close awhile, for here comes one in haste' (1.3.131). Those last six words also appear in *Much Ado About Nothing*:

BENEDICK
And how do you?
BEATRICE
Very ill too.
BENEDICK
Serve God, love me, and mend. There will I
leave you too, for here comes one in haste.

(5.2.83–6)

This six-word sequence sounds commonplace enough, but it does not appear anywhere in the fifty-one plays by the other four authors, or in the eighty-eight plays by other writers. *Julius Caesar* also shares a second rare sequence, 'I thank you for your pains' (2.4.115), with *Cymbeline* (1.6.202), *Much Ado* (2.3.240), *Taming of the Shrew* (3.2.183) and *Twelfth Night* (1.5.275), and a third, 'and bring me word what he' (2.4.47), with *1 Henry IV* (5.1.109–10), giving it a score of three for Shakespeare links (see Table 3.1.1).[4]

If we expanded our set of plays, or changed the rules in other ways, we might find examples of these sequences in other authors. But under precisely the conditions specified, we do not. Given an equal chance to emerge in parent author sets and in others, links to the parent authors dominate. We also ensure comparability by following standard protocols in preparing the texts. In all of them, spelling is modernized, for instance, contractions are expanded in the same way, and the same grammatical functions are marked. Stage directions and other text not meant to be spoken or sung on stage are excluded.

The analysis in Table 3.1.1 is limited to one feature, six-word sequences, and one way of collecting statistics about this feature, but there is no reason to think that the same pattern would not hold true for five-word sequences, four-word sequences or a different selection of plays or authors. It is supported by a large number of Shakespeare authorial studies (e.g. Jackson 2014, and Taylor and Egan 2017), which

typically begin with a test of the reliability of the scheme proposed to determine authorship, using test plays and play portions of known authorship.

Another way to test the hypothesis of authorial distinctiveness is to examine the relative frequency of individual words, focusing on the very common ones like *the* and *know*. Here we are interested in marked and consistent differences in frequencies between canons. The obvious way to assess this kind of difference would be to observe whether averages of a word are higher or lower in one author compared to another, but this would not take into account the variation from work to work. If a frequency changes wildly through a corpus, then a high or low count in a mystery text does not tell us much.

The '*t*-test', first introduced by W. S. Gosset in 1908, takes account of this element of variation in counts by dividing the difference between the two averages by the combined standard deviations of the two sets of counts (Craig and Greatley-Hirsch 2017: 50–2). 'Standard deviations' are measures of dispersion around the average. The higher the standard deviation, the greater the dispersion. Thus, the size of the final *t*-statistic is moderated by the amount of variation. The *t*-statistic has a predictable distribution, given the number of observations, so we can tell how often a given score would come about by chance alone.

We take each of 100 very common words and test how many of them are consistently different in frequency in pairs of authors. We take the same five authors as before, but this time use all available works by them, since the *t*-test results are not affected by differences in the amount of observations in the two sets being compared. We adopt the threshold of a probability of 1 in a 100 that the result could have come about by chance. A *t*-test score at that level or higher is counted as a highly significant difference. There are ten two-way comparisons possible between the five authors. Table 3.1.2 shows the results.

The shaded cells are comparisons of one playwright group with itself, or repeats of comparisons in the upper right portion of the chart. The comparisons of interest are in the unshaded cells. The lowest score recorded is seventeen, for the comparison between Chapman and Jonson. The expectation for random data is that one word out of the 100 tested would be significantly different, so even seventeen is an unexpectedly high score. The table shows strong differentiations between

Table 3.1.2 Numbers of Significant Differences in the Frequencies of 100 Very Common Words in Comparisons among Five Authors

	Chapman	Fletcher	Jonson	Middleton	Shakespeare
Chapman		35	17	32	21
Fletcher			38	32	45
Jonson				31	34
Middleton					46
Shakespeare					

Table 3.1.3 Numbers of Significant Differences in the Frequencies of 100 Very Common Words in Comparisons among Five Random Sets of Plays. ~~Numbers of Significant Differences in the Frequencies of 100 Very Common Words in Comparisons among Five Authors~~

	Random group A	Random group B	Random group C	Random group D	Random group E
Random group A		0	2	1	0
Random group B			0	0	0
Random group C				1	0
Random group D					2
Random group E					

the authors. The highest counts are forty-five for the comparison of Fletcher and Shakespeare, and forty-six for the comparison of Middleton and Shakespeare.

To check that these counts do not arise from chance local concentrations of plays of one genre, era or type, we can make up some random groups of plays and try the same test. To align with the five-author comparison set, we establish groups of different sizes, thirteen, sixteen, seventeen, nineteen and twenty-eight, to match the sizes of the Chapman, Fletcher, Jonson, Middleton and Shakespeare sets. Table 3.1.3 shows the results.

In six of the comparisons, there are no significantly different words, in two of them there is one, and in two of them there are two. This indicates that the test of word frequency profiles in groups of plays is working as expected, and that the theoretical expectation of one significantly different word in randomly mixed sets is broadly confirmed in practice. Chance throws up the occasional significant difference, but no more than that.

As already mentioned, the second largest number of significant differences in the author comparisons is in the comparison of Fletcher and Shakespeare (Table 3.1.2). These two playwrights collaborated on plays and are thus a natural pair for investigation. Of the forty-five words the largest *t*-test score and thus the most significant difference is for rates of use of *in* as a preposition ('she is in the court' rather than 'she went in just now'). Figure 3.1.1 shows percentage counts for Shakespeare plays to the left and for Fletcher plays to the right.

The highest percentage count of this word among the Fletcher plays, ~~at the right-hand end of the chart,~~ is in *Rule a Wife and Have a Wife*, 0.9 per cent, but this is still lower than the lowest count for a Shakespeare play, ~~at the left hand end of the chart,~~ just under 1 per cent, in *The Winter's Tale*. Thus, the frequency of this word provides a complete separation between Fletcher and Shakespeare plays. This is true also of one other word in the set of one hundred, *that* as a conjunction, where the lowest Shakespeare score is again higher than the highest Fletcher score.

We can conclude that there are persistent internal consistencies in authorial canons. In both our examples, shared rare sequences and different rates of use of very common words, we gave the test of authorial consistency a chance to fail. The

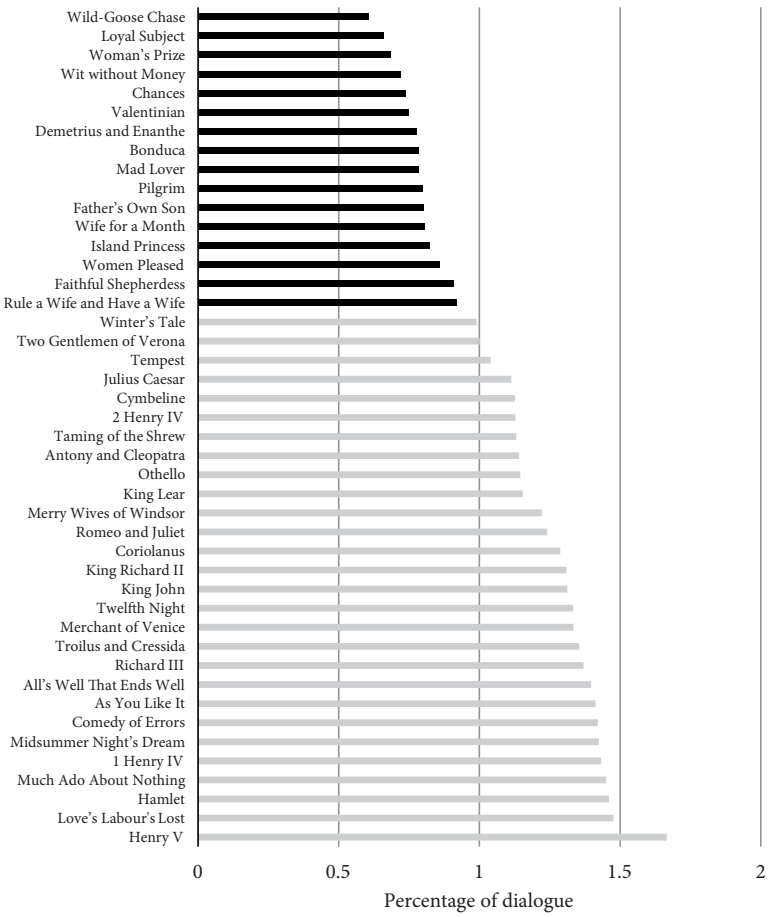

Figure 3.1.1 Percentage Counts of *in* as a Preposition in 28 Plays by Shakespeare and 16 Plays by Fletcher.

rare sequences could have appeared across the comparisons with no real pattern. There could have been no significant differences in frequency between canons, or differences appearing equally in canons and random assemblages. Canons, when given the chance, behave like clusters of works with marked strands of similarity. This authorial distinctiveness and self-consistency is the foundation for work in attribution. It is never to be taken for granted, since it does not necessarily appear in every mode, genre, period, and there are questions, to which we turn below, about exceptions, and small samples and small canons. But the two studies above have shown that it is reasonable to start a Shakespeare attribution study with an assumption of persistent underlying authorial difference.

Hath, the older form of *has*, is another common word which Shakespeare and Fletcher use at different rates, as the *t*-test confirms. Shakespeare's average is much higher. There is one stark exception to the Fletcher pattern, however (see Figure 3.1.2).

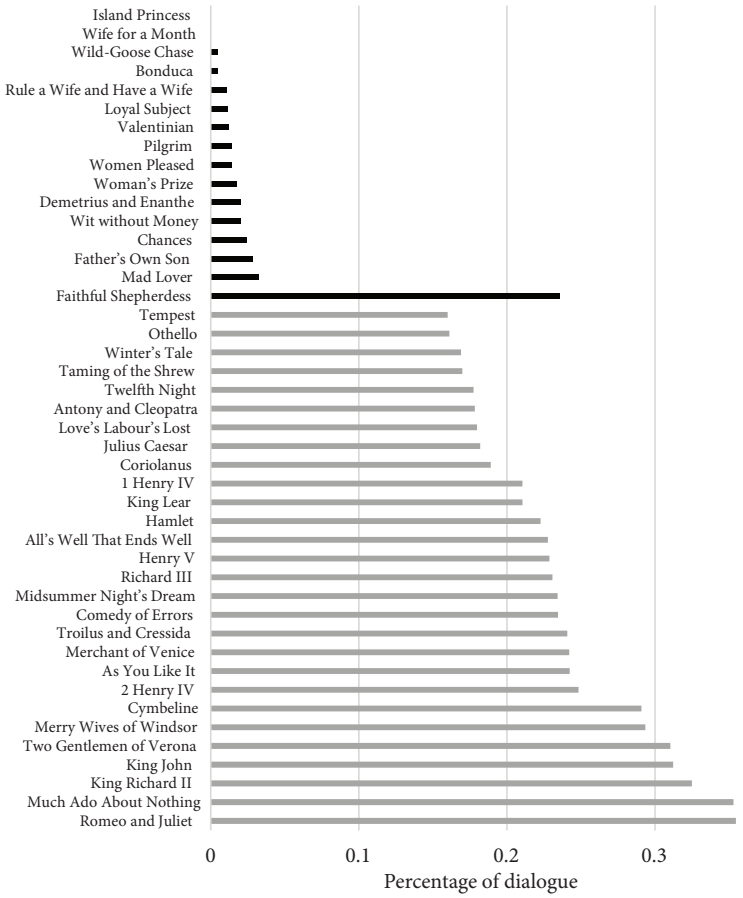

Figure 3.1.2 Percentage Counts of *hath* in 28 Plays by Shakespeare and 16 Plays by Fletcher.

Fletcher counts are generally much lower than Shakespeare's, and two Fletcher plays have no instances at all, but *The Faithful Shepherdess* has forty-seven instances, 0.24 per cent of its dialogue. The next highest, *The Mad Lover*, has six instances, or 0.03 per cent. Fletcher avoids *hath* generally, in strong contrast to his sometime collaborator Shakespeare, but departs from his general practice to a marked degree on one occasion. (There is always the possibility that in this case, exceptionally, a scribe or compositor altered the forms they found in their copy, but equally there is nothing in the bibliography of *Faithful Shepherdess* that indicates this.) If the authorship of this play was disputed, and we were relying on this marker, it would seem to offer strong evidence that Shakespeare is a more likely author than Fletcher. Cyrus Hoy and Jonathan Hope both excluded *Faithful Shepherdess* from their Fletcher reference sets for attribution purposes (Hoy 1956; Hope 1994). At times authors can confound standard authorship methods and write unlike themselves throughout a work.

We have to reckon also with the question of smaller samples. The texts we are interested in are often acts, scenes or even parts of scenes rather than whole plays. These shorter samples are inherently more likely to diverge from a general authorial standard, according to the statistical law of large numbers. Larger samples give more reliable results because more data lies behind them. To illustrate the importance of sample size, we can take the case of characters' spoken parts in plays. In Shakespeare's core canon the number of words spoken by a character (excepting 'mute' characters) ranges from one – 'Aye' (the second senator in *Cymbeline*) or 'Stand' (the thieves in *1 Henry IV*) – to Hamlet's 11,328. Usually we express the frequency of a particular word as a percentage, to take account of variations in size like this, but this may disguise the difference in meaningfulness between the percentage score in a small character part and the percentage score in a large one. Figure 3.1.3 shows the percentages of the word *the*, generally the commonest word in plays, for the 578 characters with 100 words or more in the core Shakespeare canon.

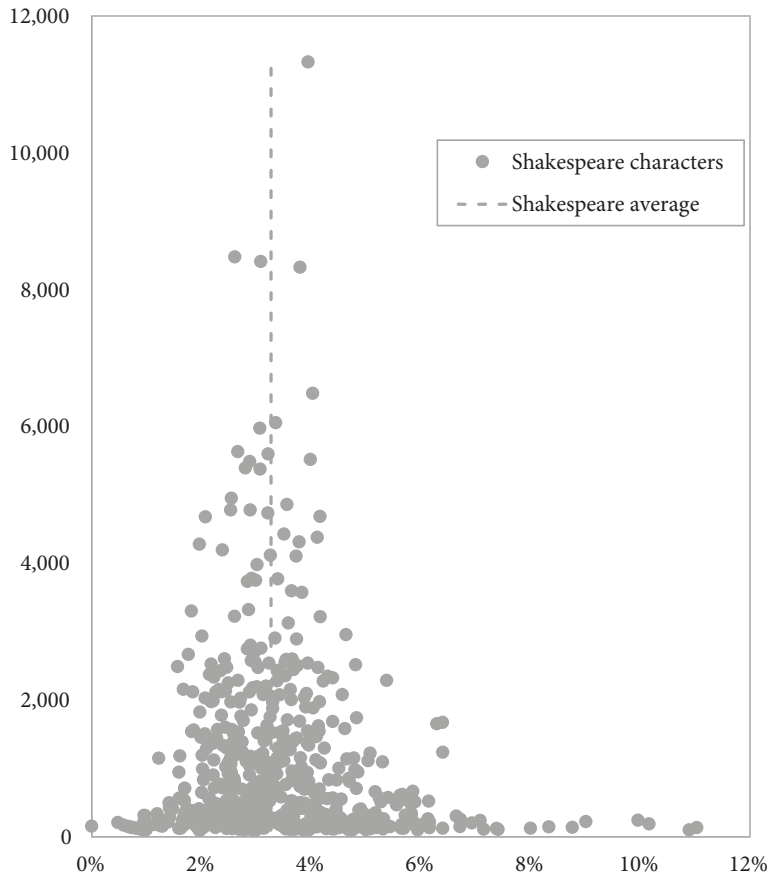

Figure 3.1.3 Percentages of the Word *the* in Larger Shakespeare Character Parts, Versus Size of Part.

Hamlet is at the top, with more than 11,000 words, as already noted. The datapoints for smaller characters are at the bottom of the chart. Their scores are widely scattered across the horizontal axis, which represents the percentages of dialogue represented by *the*. Mamillius in *The Winter's Tale* is at the left-hand extreme. He speaks 155 words without using *the* at all, so has a score of 0 per cent. At the other extreme is the Second Gentleman in *Othello*, who speaks 136 words, of which fifteen are *the*, a score of 11 per cent. The larger characters cluster much more closely around the overall Shakespeare average, which is 3.3 per cent (20,418 instances of *the* in 625,041 words in the twenty-eight plays). Hamlet, at the top, has 4 per cent, then in descending order come Iago, Richard III and Henry V, with 2.6 per cent, 3 per cent and 3.8 per cent, respectively. With the larger parts, one-off factors are balanced out and underlying consistent trends win out. The difference between the percentages of Hamlet's and Iago's parts for the chosen word, 4 per cent for Hamlet, and 2.6 per cent for Iago, are more meaningful than this same difference in characters with very small parts, where we could just about dismiss the variation as a chance effect.

In authorship attribution, percentage use of words like *the* is often an important marker. Shakespeare's average is 3.3 per cent, as already noted, whereas Fletcher's is much lower, 2.4 per cent (8,072 instances in a canon of sixteen plays totalling 340,719 words). We would expect 621 instances of *the* in an average 19,000-word play by Shakespeare, but 450 instances in a play of the same length by Fletcher.[5] Yet Shakespeare-like or Fletcher-like percentages mean little in small samples, because they may be simply the result of chance. In Figure 3.1.4 the scores for Fletcher characters for *the* are plotted and shown with the scores for Shakespeare characters repeated from Figure 3.1.3.

As with the Shakespeare characters, and again following the law of large numbers, as the Fletcher character parts get larger, they disperse less widely around the overall Fletcher average. (One black circle spoils the neat Christmas tree shape made by the Fletcher entries, with a count of *the* much higher than the Fletcher average despite a relatively large size of part. This is the Satyr from *Faithful Shepherdess*, with 1,931 words in all, of which five and a half percent are *the*. Fletcher's pastoral is once again anomalous.) The tendency for Shakespeare characters to use more *the* proportionally than Fletcher characters is clear overall, but most of the smaller characters, those below 2,000 words, are in shared territory, with more overlap than distinctiveness. Given the choice, one would always want a larger sample with its consequent greater reliability. A percentage score is more weighty for purposes like attribution when based on a large sample.

The same rule applies to authorial canons in attribution. We can think of the process as one of prediction. We have a set of sole-authored works by a given author, and we aim to use the patterns found there to predict how an author would write in their next attempt in a similar text type. If, as in the case of Thomas Nashe, there is just one sole-authored play available, we are on shaky ground. This work, *Summer's Last Will and Testament*, is a comedy, but we need a guide to Nashe's writing more broadly, if we are investigating his possible contribution to the history play *1 Henry VI*, or the tragedy *Dido, Queen of Carthage*. Christopher Marlowe also has a claim to *Dido*, but his canon of six plays for comparison quickly shrinks

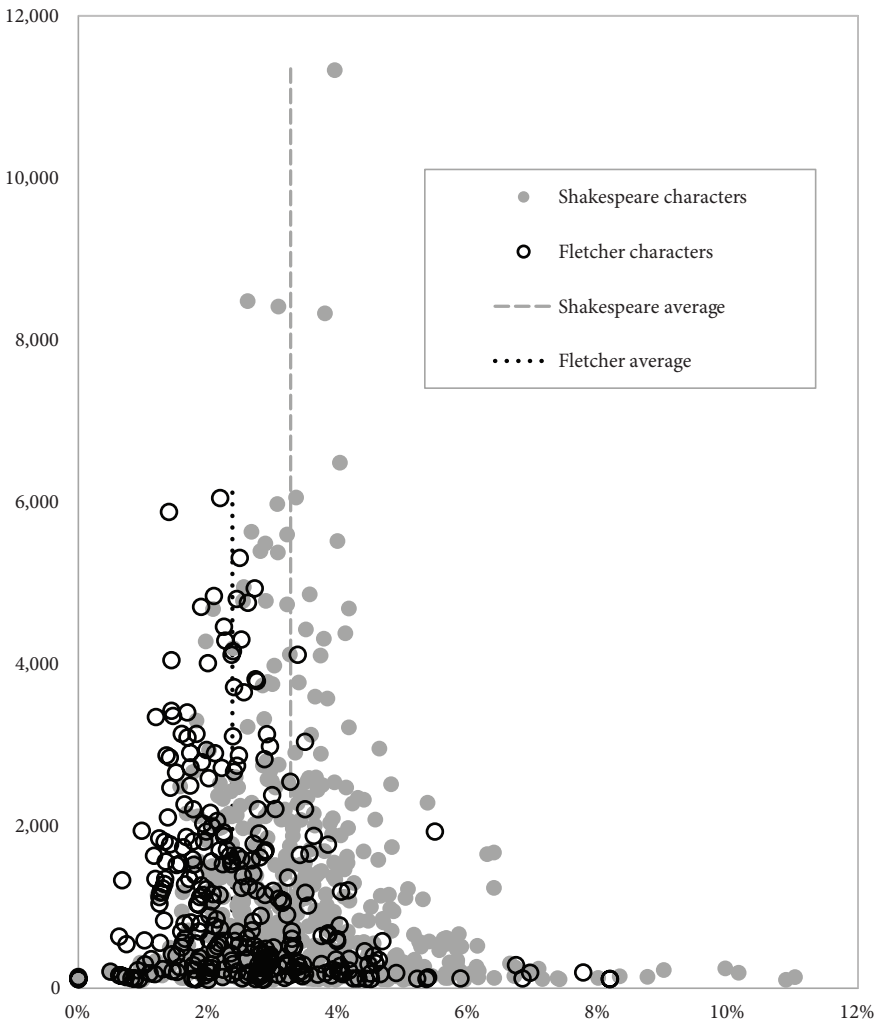

Figure 3.1.4 Percentages of the Word *the* in Larger Shakespeare and Fletcher Character Parts, Versus Size of Part.

on examination. *Doctor Faustus* is very likely a collaboration, and exists in two different early editions, one much shorter than the other; *The Jew of Malta* may well have been revised by another author; and the surviving version of *The Massacre at Paris*, has evidently suffered in transmission (see the entries for the various plays in Wiggins and Richardson 2012–). Thomas Kyd is another author who is often a person of interest in attributing anonymous and disputed plays, but just one well-attributed original play, *The Spanish Tragedy*, is available as the basis to judge the potential stylistic range and limits of his writing, though we may be tempted to relax the requirements and add a translation, *Cornelia*, and one further anonymous play, *Soliman and Perseda*, which is very likely to be his.

By contrast Shakespeare is a haven of safety with twenty-eight well-attributed single-author plays. This is larger than any other surviving early modern dramatic canon apart from that of James Shirley. The canons of Chapman, Fletcher, Jonson and Middleton, already discussed in connection with questions of authorial consistency, also qualify as beneficiaries of the law of large numbers. We can be more confident in attributing plays and play sections between any of these five than with Nashe, Marlowe and Kyd.

So far we have concentrated on single variables like the very common words or, as with the six-word sequences, on a simple accumulation of instances in a category. Readers can check a particular concentration or dearth against their own perceptions of the language of a work. A given number of concrete occurrences is a secure and simple foundation, introducing the minimum of potentially sophisticating transformations, and only going as far as expressing the raw counts as proportions to take account of different sample sizes.

There is also a battery of 'multivariate' classification procedures which use a number of variables together (Tabachnick and Fidell 2018). They combine the separate discriminatory power of single variables into a composite method with power to classify in more difficult cases. One example of a procedure of this kind is Linear Discriminant Analysis. This is based on a method invented by Sir Ronald Fisher and first announced in 1936. We concentrate here on its simplest form in which there are just two classes to be considered – for instance, Author A and Author B, or Author A and a composite set of other authors. Each of the variables is given a weighting, high or low, positive or negative, with the objective of building a composite function which does the best possible job of delivering scores which divide the two classes. The ideal outcome is a function on which all the texts in one class have higher scores, or all have lower scores, than all the texts in the other class.

We can consider applying this method to a standard quantitative authorship attribution project which has a mystery text and a pair of candidate authors. The first step is to seek out the distinctive characteristics of the candidate authors, and the second is to compare these profiles of features with the patterns in the mystery text. When this process is automated, we can speak of making a 'classifier' and 'training' it on samples of text known to be by the candidate authors. The procedure works through the data it is given, to find variables on which the authors differ consistently and to compile them into a single test.

Now a second group of considerations emerge. The purpose of the exercise is to make the correct assignation of the mystery text to one class or the other. To prepare for this we have trained the classifier on the available texts of known provenance and the available features, and our measure of success is performance with those. However, we have as yet no way of knowing how well this classifier will perform with a freshly introduced text. We may have a classifier which is perfectly adjusted to the training set but does not generalize well to unknowns. This is the problem of 'overfitting' – tailoring a classifier to a training set while ignoring the ultimate purpose of the exercise, which is dealing with new items.

The usual approach to this problem of estimating reliability with freshly introduced samples is to use some of the training set as 'test' items. We take some

samples out of the training set at the very beginning, train the classifier without them and then use the classifier to assign the reserved samples to one class or the other. In this case we know the true class to which the sample belongs, so this is a good check on the reliability of the classifier.

To demonstrate the workings of these 'training' and 'test' sets we can carry out a Linear Discriminant Analysis which does not focus on a mystery text, but rather concentrates on evaluating the reliability of the method. We use the core canon of twenty-eight Shakespeare plays as one class and a set of 138 single-author well-attributed plays by others, with dates of first performance 1580–1619, as the second. We divide all the plays into 2,000-word segments. This yields 301 Shakespeare segments and 1,253 non-Shakespeare ones. We have 1,000 word-variables available, percentages of 1,000 very common words, having counted instances of these words in all the segments. With the assistance of the statistics program SPSS we create a Linear Discriminant Function maximising the distinction between the two classes (IBM Corp. 2017).[6] This achieves a perfect separation of the Shakespeare and non-Shakespeare classes.

To provide a guide to the severity or otherwise of overfitting in a given case, SPSS offers one standard form of 'test' sampling, known as 'cross-validation'. A single segment from the training sets is withheld at the beginning and classified at the end, the result (right or wrong) is noted, then this sample is replaced and another is reserved and classified, and so on, until all segments have been used in this way. The program reports that in the end 261 out of 301 Shakespeare test segments and 1,158 out of 1,253 non-Shakespeare test segments were correctly assigned, or in all 1,419 out of 1,554 segments, 91.3 per cent. This is much lower than the result for the training segments, which was 100 per cent, but we could regard it as more accurate as a prediction of what would happen with a mystery text, since it comes closer to emulating the circumstances of the classification of a newly introduced sample. These newly introduced samples are likely to have idiosyncrasies which are not taken account of in the process of building the classifier, whereas the anomalies in the members of the training sets (on both sides) are accommodated in arriving at the weightings in the particular function which emerges after the training process.

This result is probably still flattering to the classifier, however, since while each segment is new to the particular classifier which is built without it, the remaining segments of the longer text from which the test segment comes are not. They remain in the relevant training set. We can expect considerable likeness within the segments of a play, for instance, much greater than the likeness between the segments of different plays, so some of the idiosyncrasies in the cross-validation segment will have been taken account of this way in the classifier. This would not be true of a mystery set of segments. For this reason, we perform a second test of the test, this time reserving a group of whole plays, with all their segments, at the beginning, and classifying them at the end once the classifier has been trained on the remaining plays and their segments. We choose a quarter of the Shakespeare plays and a quarter of the non-Shakespeare plays at random and make these the test set while the training set is the remaining three-quarters of the two original groups (Table 3.1.4).

This time we find that 319 out of a total of 389 segments have been correctly assigned, or 82 per cent. This test set result is much lower than for the training set

Table 3.1.4 Training and Test Sets

	Plays	Segments
Shakespeare training	21	230
Shakespeare test	7	71
Non-Shakespeare training	104	935
Non-Shakespeare test	34	318
Total	166	1554

(100 per cent). Given this difference we can say the function is 'overfitted' to the training set, i.e. it is wonderfully well attuned to the variation in the training set, but so much so that its performance with a test set is relatively poor.

We have the option of using fewer variables, in an attempt to moderate this overfitting aspect. Fewer variables means less opportunity to fit the classifier minutely to the training set. Figure 3.1.5 shows what happens when we use first only the 100 most common words, then the 200 most common words, and so on, moving in 100-word jumps up to the set of 1,000 we began with.

Working backwards from the result we have already discussed, using 1,000 words, shown here at the right-hand end of the chart, we can see that the performance of the classifier in the cross-validation by segment and in the whole-play test is much better with fewer words, peaking at 400 words for both (at 94.3 per cent and 92.3 per cent, respectively). At 100 words the performance of the classifier is a little better with the whole-plays test than with the cross-validation set, but thereafter the performance with the test set is lower, and by a generally increasing margin. At 400 words, presumably, there is the advantage of having more markers, and a better chance of including good individual markers, than with the earlier marker sets, but the problem of overfitting has not yet set in. The performance with the training set reaches 100 per cent with 600 words and this is maintained through the rest of the trials. The descent of the dashed line and the tramline after 400 words visualizes the overfitting that is taking place. As the procedure includes more words and chooses weightings for them to maximize success with the training set, the function performs less and less well with freshly introduced segments.

There is no way of knowing how closely the pattern of Figure 3.1.5 would be repeated with a different authorial contrast, smaller or larger segments, a different feature set or a different statistical procedure. It would be risky to rely too much on the fact that in this case the peak is at 400 words. On the other hand, this demonstration does serve to illustrate some general relationships that will emerge regularly in work of this kind. More markers make for a better performance up to a point, and then they bring the risk of overfitting; cross-validation performance counts will generally be lower than training set results; and whole-text test results will generally be lower again. We can note in passing that at its best, in a whole-text test set, this system correctly assigns 2,000-word segments to Shakespeare, and away from him, ninety-two times out of a hundred. This suggests we can be confident of detecting a powerful author effect in studies like this, whatever the sceptics may say, but we also have to remember that we can be sure of making some errors in the attributions.

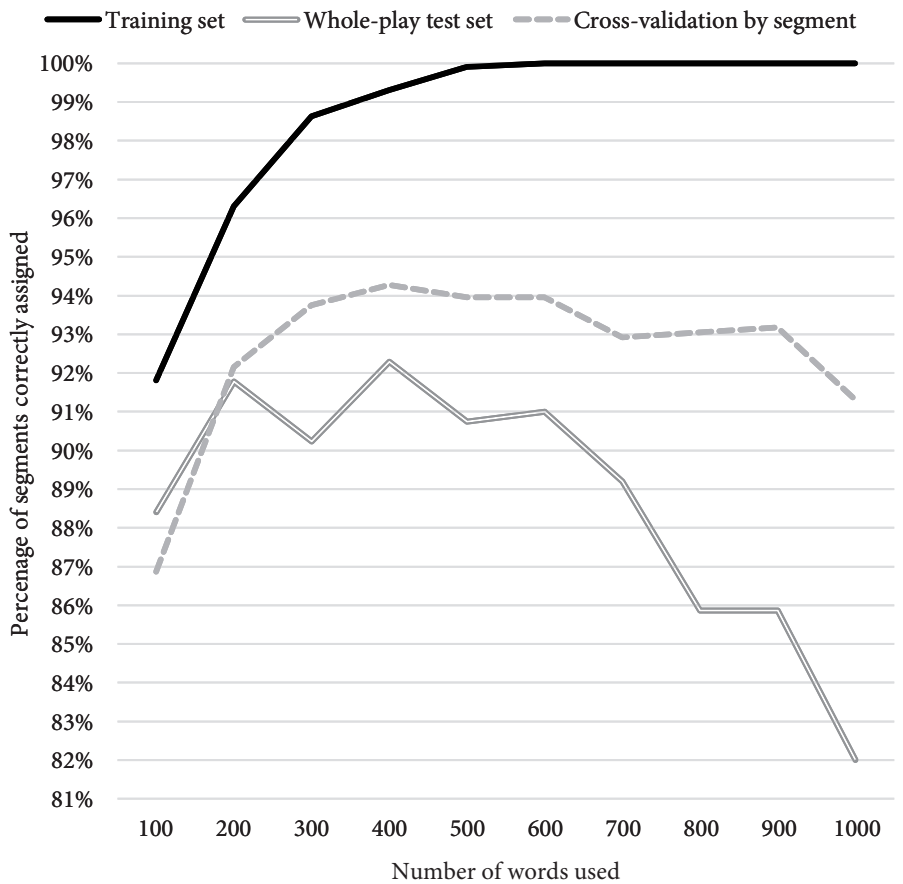

Figure 3.1.5 Success Rates for the Classification of Shakespeare Plays and Non-Shakespeare plays Using Linear Discriminant Analysis and Ten Different Word-variable Sets.

My aim in this chapter has been to establish some principles to guide the scholar embarking on authorship attribution work. I hope it is now evident that they can have confidence that an author effect exists, but must be aware that this effect is not necessarily even, given the myriad forces making for variation in style. They can be encouraged by the power of multivariate tests, but know that this power brings its own problems. Once they have a method and a corpus which seems adequate, they must always return to the fundamental question: how often does the new classifier assign correctly samples as close as possible to the mystery text in size and type, but of known provenance? The saving grace of quantitative authorship attribution is that so much of it is testable. We do not have to rely on *a priori* axioms, *ex cathedra* pronouncements or common sense beliefs about the strength or otherwise of a given kind of evidence or given method, but can estimate this strength for ourselves and watch the fascinating contest between pattern and variation in literary language play out.

PRACTICALITIES

In an ideal world, Shakespeare attribution work would be supported by open data and open tools, so that steady progress could be made as methods are compared and experimental design is refined, but this is some way off. Fortunately, some open-source, well-supported tools do exist. *Stylo*, for instance, is a package for the statistical computing environment R, which is tailor-made for stylometry (Eder, Rybicki and Kestemont 2016). It can take the neophyte researcher from raw text to statistical procedures. Good metadata is also available. The online DEEP: Database of Early English Playbooks is limited only by the exclusion of manuscript plays from its remit (Farmer and Lesser 2007). Researchers can download the entire set of items and attributes, or search the information online. Nine of ten volumes of the *British Drama 1533–1642: A Catalogue* have now been published (Wiggins and Richardson 2012–). This gives a wonderful amount of detail – down to the props used in a given play – and offers the most authoritative current judgement across the board of questions like authorship and date. Metadata is a critical aspect of authorship study. With its help, the attributionist can strengthen a corpus for a given purpose by excluding some works, like translations, plays written for reading rather than performance, and masques and entertainments, or can test to see whether a fancied association such as a frequency and date is statistically significant.

It is disappointing to report that there is no well-edited comprehensive corpus of Shakespeare-era drama supported by an enduring institution and free to download. Almost all printed texts from the period now exist as digitized images, but machine-readable text requires transcription by mechanical means like Optical Character Recognition or human means like keyboarding, and then a further stage of editing. Quality edited machine-readable texts based on individual early editions or manuscripts are still in short supply. The situation for Shakespeare texts, as ever, is the exception. Internet Shakespeare Editions (Best 1996-) and the Oxford Text Archive (n.d.) both make available proofread texts based on individual early printed versions. For texts beyond Shakespeare, EEBO-TCP offers a vast array of titles, but they contain many gaps where keyboarders found the text hard to read (Early English Books Online Text Encoding Partnership 2015). Literature Online, or 'LION', has a large set of well-edited texts, but these are for searching rather than downloading, and they require a subscription for access (ProQuest 2019).

New Shakespeare attributionists will therefore need to find texts wherever they can. They will most likely have to do some editing, to fill in gaps or to convert old spelling to standard modern and to expand contractions. They may also want to lemmatize, collecting different forms (such as *cry*, *cries*, *cried* and *crying*) under a single dictionary headword, and they may want to tag words for parts of speech so as to distinguish various homographs (such as *that* in 'she said that she would,' 'see that sword', and 'the book that I left'). Software exists to help with these tasks, such as VARD 2 for modernising spelling (Baron 2013), but none of it is capable of a good result on early modern English text without human input, so there is considerable labour involved, and the larger the corpus and the higher the standards of accuracy, the more laborious it becomes.

Texts for counting should be prepared in a consistent way, and there is much to be said for separating the text proper from other material like prefaces, dedications, commendatory verse and footnotes, and, in the case of drama, from speaker prefixes and stage directions as well. Otherwise instances of *all* in footnotes may be inadvertently included in totals for a poem (Craig 2012b: 171n.58); a repeated speaker tag like 'An.' may inflate the totals for the word *an* in dramatic dialogue (Jackson 1999); or counts of *exeunt* may be mistakenly added to a list of authorial markers derived from a corpus mixing drama and non-dramatic materials (Freebury-Jones and Dahl 2019: 5). The Text Encoding Initiative offers a comprehensive standard set of tags to encode this parsing of the text (Text Encoding Consortium 2019).

Before plunging into a study, a researcher new to the field might consult some background work on authorship and attribution (Love 2002; Craig 2012a) and read some model studies (Vickers 2002; Jackson 2014; Taylor and Egan 2017). Shakespeare attribution is not for the faint-hearted. The required investment in text preparation and in learning methods is considerable. Any findings will be given intense and sometimes hostile scrutiny before and after publication, since the stakes are high and positions on the many contested questions are entrenched. On the other hand, the intersection of statistics and language, and its application to works of towering cultural prestige, makes for a heady mix, and I imagine this will continue to prove irresistible for those of a certain temperament, though sometimes against their better judgement.

NOTES

1. Thanks to Ruth Lunney and Brett Greatley-Hirsch for very helpful comments on an earlier version of this chapter.

2. The lists of plays are in the 'Supplementary Materials' for this chapter, to be found at http://hdl.handle.net/1959.13/1406580, where the reader will also find extra background data and metadata to fill out details of each of the tables and charts in the chapter.

3. Strictly speaking, the playing field is not entirely level because the five chosen dramatists have different size canons in the corpus – Chapman thirteen, Fletcher eleven, Jonson twelve, Middleton fifteen and Shakespeare twenty-eight. This means that the comparison set for each playwright varies in size. This does not affect comparisons up and down the columns of Table 3.1.1, since the ten test plays in each column have exactly the same comparison set.

4. We are counting the number of different sequences that are repeated, rather than tallying up all the instances of these sequences, so these sequences count as one link for the purposes of Table 3.1.1.

5. In the whole-plays comparison between Shakespeare and Fletcher, *the* has a *t*-test result which corresponds to a probability of 0.0000005 of coming about by chance.

6. In SPSS, go to 'Analyze', then 'Classify' and finally 'Discriminant'.

REFERENCES

Baron, Alistair (2013), *Variant Detector (VARD) 2*, Leicester: University of Leicester.

Best, Michael, coordinating ed. (1996-), *Internet Shakespeare Editions*. Available online: https://internetshakespeare.uvic.ca/(accessed 3 October 2019).

Craig, Hugh (2012a), 'Authorship', in Arthur F. Kinney (ed.), *The Oxford Handbook of Shakespeare*, 15–30, Oxford: Oxford University Press.

Craig, Hugh (2012b), 'George Chapman, John Davies of Hereford, William Shakespeare, and *A Lover's Complaint*,' *Shakespeare Quarterly*, 63 (2): 147–74.

Craig, Hugh and Brett Greatley-Hirsch (2017), *Style, Computers, and Early Modern Drama*, Cambridge: Cambridge University Press.

Early English Books Online Text Encoding Partnership (2015), *EEBO-TCP Phase 1*, Oxford and Ann Arbor, MI.

Eder, Maciej, Jan Rybicki and Mike Kestemont (2016), 'Stylometry with R: A Package for Computational Text Analysis', *R Journal*, 8 (1): 107–21.

Farmer, Alan B. and Zachary Lesser (2007), DEEP: Database of Early English Playbooks. Available online: http://deep.sas.upenn.edu (accessed 3 October 2019).

Fisher, Ronald A. (1936), 'The Use of Multiple Measurements in Taxonomic Problems', *Annals of Eugenics*, 7 (2): 179–88.

Freebury-Jones, Darren and Marcus Dahl (2019), 'Searching for Thomas Nashe in *Dido, Queen of Carthage*', *Digital Scholarship in the Humanities*. Available online: https://doi.org/10.1093/llc/fqz008 (accessed 3 October 2019).

Gosset, William S. [writing as 'Student'] (1908), 'The Probable Error of a Mean', *Biometrika*, 6 (1): 1–25.

Hope, Jonathan (1994), *The Authorship of Shakespeare's Plays: A Socio-Linguistic Study*, Cambridge: Cambridge University Press.

Hoy, Cyrus (1956), 'The Shares of Fletcher and his Collaborators in the Beaumont and Fletcher Canon (I)', *Studies in Bibliography*, 8: 129–46.

Hunter, George K. (1997), *English Drama 1586–1642: The Age of Shakespeare*, Oxford: Clarendon Press.

IBM Corp (2017), *IBM SPSS Statistics for Windows*, Version 25.0, Armonk, NY: IBM Corp.

Jackson, MacDonald P. (1999), '*Titus Andronicus* and Electronic Databases: A Correction and a Warning', *Notes and Queries*, 244 (NS 46.2): 209–10.

Jackson, MacDonald P. (2014), *Determining the Shakespeare Canon: 'Arden of Faversham' and 'A Lover's Complaint'*, Oxford: Oxford University Press.

Love, Harold (2002), *Attributing Authorship: An Introduction*, Cambridge: Cambridge University Press.

McMullan, Gordon, ed. (2000), *King Henry VIII (All is True)*, The Arden Shakespeare, London: Thomson Learning.

Oxford Text Archive (n.d.). Available online: https://ota.ox.ac.uk (accessed 3 October 2019).

ProQuest (2019), Literature Online, Ann Arbor, Michigan: ProQuest, Inc. Available online: https://about.proquest.com/products-services/literature_online.html (accessed 30 September 2020).

Rudman, Joseph (2016), 'Non-traditional Authorship Attribution Studies of William Shakespeare's Canon: Some Caveats,' *Journal of Early Modern Studies*, 5: 307–28.
Tabachnick, Barbara G. and Linda S. Fidell (2018), *Using Multivariate Statistics*, 7th edn, Harlow, Essex: Pearson Higher Education.
Taylor, Gary and Gabriel Egan, eds (2017), *The New Oxford Shakespeare: Authorship Companion*, Oxford: Oxford University Press.
Text Encoding Consortium (2019), *TEI P5: Guidelines for Electronic Text Encoding and Interchange, 3.6.0*. Available online: http://www.tei-c.org/Guidelines/P5/ (accessed 3 October 2019).
Vickers, Brian (2002), *Shakespeare, Co-Author: A Historical Study of the Five Collaborative Plays*, Oxford: Oxford University Press.
Wiggins, Martin and Catherine Richardson, eds (2012–), *British Drama 1533–1642: A Catalogue*, 10 vols, Oxford: Oxford University Press.