CHAPTER 1

Methods

In this chapter, we outline our principles of text selection and preparation and then describe the statistical and computational methods we employ throughout this book. Each description includes a working example to demonstrate the method.

Text Selection and Preparation

Appendix A lists the full-text corpus of plays we use throughout this book, along with their authors (where known), dates of first performance, the source text we use, its date of publication, and its genre. We depart from our main bibliographical source, the second edition of the *Annals of English Drama, 975–1700* (hereafter '*Annals*'), only where new research is persuasive and sound, as with the attribution of *Soliman and Perseda* to Thomas Kyd.[1]

To construct our corpus of machine-readable (that is, electronic) texts, we have relied upon base transcriptions from *Literature Online*, checked and corrected against facsimiles from *Early English Books Online*. Since our analysis concerns word frequency and distribution, and not orthography, spelling was regularised and modernised. For the sub-set of plays used in Chapter 6, this was done using VARD, a software tool developed by Alistair Baron for regularising variant spelling in historical corpora.[2] Spelling was

[1] Alfred Harbage and Samuel Schoenbaum, *Annals of English Drama, 975–1700*, 2nd edn (Philadelphia: University of Pennsylvania Press, 1964). The authors also consulted available volumes of Martin Wiggins's (in association with Catherine Richardson) *British Drama, 1533–1642: A Catalogue*, 10 vols. (Oxford University Press, 2011–), and Alan B. Farmer and Zachary Lesser (eds.), *DEEP: Database of Early English Playbooks* (2007–). On Kyd's authorship of *Soliman and Perseda*, see Lukas Erne, *Beyond 'The Spanish Tragedy': A Study of the Works of Thomas Kyd* (Manchester University Press, 2001), 157–67, as well as his Introduction to the Malone Society Reprints edition of the play (Thomas Kyd, *Soliman and Perseda*, ed. Lukas Erne (Manchester University Press, 2014)).

[2] See Alistair Baron, Paul Rayson, and Dawn Archer, 'Word Frequency and Key Word Statistics in Historical Corpus Linguistics', *Anglistik* 20.1 (2009), 41–67. While *VARD* can be trained to regularise words algorithmically (i.e., when a given certainty threshold is met) with little to no human supervision or intervention, we instead used *VARD* as a tool to generate a list of variant word types

https://doi.org/10.1017/9781108120456.002 Published online by Cambridge University Press

modernised, but early modern English word forms with present tense *-eth* and *-est* verb-endings (e.g. *liveth* and *farest*) were retained. For the larger sets of plays utilised in Chapters 2 and 7, we regularised spelling using a function in the Intelligent Archive software to combine variant forms with their headwords.[3] In all texts used, function words with homograph forms – such as the noun and verb form of *will* – were tagged to enable distinct counts for each. Appendix E lists the function words used in our analysis. Contractions were also expanded, such that where appropriate 'Ile' was expanded as an instance of *I* and one of $will_{\text{verb}}$, 'thats' as an instance of $that_{\text{demonstrative}}$ and one of *is*, and so on.

Unless otherwise specified, texts are segmented into non-overlapping blocks of words – typically 2,000 words – with the last block, if incomplete, discarded to ensure consistent proportions. Proper names, passages in foreign languages, and stage directions are also discarded.[4] It is standard practice in authorship attribution testing to exclude proper names and foreign-language words from the analysis, because these are more closely related to local, play-specific contexts rather than indicative of any consistent stylistic pattern. As for stage directions, Paul Werstine has demonstrated that their status as authorial or non-authorial cannot be assumed, but varies from text to text.[5] We deemed it safer to exclude stage directions as a general rule rather than attempt to assess every instance.

Principal Components Analysis

Principal Components Analysis, or PCA, is a statistical procedure used to explain as much of the total variation in a dataset with as few variables as possible. This is accomplished by condensing multiple variables that are correlated with one another,[6] but largely independent of others, into a

and provide a list of possible modern equivalents. We chose equivalents on a case-by-case basis in light of the context in which the variant spelling forms appeared.

[3] For a fuller discussion of this functionality, see Hugh Craig and R. Whipp, 'Old Spellings, New Methods: Automated Procedures for Indeterminate Linguistic Data', *Literary and Linguistic Computing* 25.1 (2010), 37–52.

[4] That is, single words in languages other than English are included, but passages with two or more consecutive words in a foreign language are excluded.

[5] Paul Werstine, *Early Modern Playhouse Manuscripts and the Editing of Shakespeare* (Cambridge University Press, 2013), esp. 123–30, 157–84.

[6] As mentioned in the previous chapter, the term 'correlation' is used in statistics to describe and measure the strength (low to high) and direction (positive or negative) of the association between two sets of counts. Counts increasing or decreasing in parallel with one another are said to have a positive correlation; by contrast, a negative correlation arises where one count increases while the other decreases (and *vice versa*).

https://doi.org/10.1017/9781108120456.002 Published online by Cambridge University Press

smaller number of composite 'factors'.[7] The strongest factor or 'principal component' is the one that accounts for the largest proportion of the total variance in the data. PCA produces the strongest factor (the 'first principal component'), and then the factor that accounts for the greatest proportion of the remaining variance while also satisfying the condition that it is uncorrelated with the first principal component – a property which we can visualise in a two-dimensional example as being at right-angles to it. Since each principal component only ever represents a proportion of the underlying relationships between the variables, PCA is a data reduction method. The method is also considered 'unsupervised', because it does not rely upon any human pre-processing of the data – the algorithm treats all of the samples equally and indifferently.[8]

A classic example of how PCA is used to reduce the dimensions of multivariate data involves taking a table of the heights and weights of a group of people from which a new composite factor – which we might call 'size' – is generated as the sum of the two variables.[9] 'Size' will represent the patterns of variation within the two original variables with a high proportion of accuracy – shorter people will tend to be lighter, and taller people heavier – but it will not account for all the possible variations in height and weight, since some short people will be heavy and some taller people light. As a principal component, 'size' still captures a basic fact about the relationship between height and weight, one that, in a sense, is the most important. If we add two variables, say, waist size and muscle mass, a new first principal component may be calculated to account for the strongest correlation between all four variables, on the same principle of accounting for most of the variation by weighing the best-coordinated variables similarly. In this scenario, waist size and weight together may represent a proxy for 'obesity', and muscle mass and weight together may represent a proxy for 'muscularity', and so on.

[7] Christopher Chatfield and Alexander J. Collins, *Introduction to Multivariate Analysis* (New York: Chapman & Hall, 1980), 57–79; and I. T. Jolliffe, *Principal Component Analysis* (New York: Springer, 1986). For a gentler introduction to the procedure, see Mick Alt, *Exploring Hyperspace: A Non-Mathematical Explanation of Multivariate Analysis* (Maidenhead: McGraw-Hill, 1990), 48–80.

[8] This is not to conflate 'unsupervised' with 'objective', as James E. Dobson rightly cautions in 'Can an Algorithm Be Disturbed?: Machine Learning, Intrinsic Criticism, and the Digital Humanities', *College Literature* 42.4 (2015), 543–64. However principled they may be, the processes of selecting and preparing the underlying corpus (outlined earlier in this chapter) are not free of subjectivity, just as all so-called 'unsupervised' methods contain human elements.

[9] As the name suggests, 'multivariate' data involves two or more variables, as opposed to 'univariate' data, which involves only a single variable.

https://doi.org/10.1017/9781108120456.002 Published online by Cambridge University Press

Table 1.1 *A select corpus of plays*

Author	Play	Date
Lyly, John	*Campaspe*	1583
Lyly, John	*Endymion*	1588
Lyly, John	*Galatea*	1585
Lyly, John	*Mother Bombie*	1591
Marlowe, Christopher	*1 Tamburlaine the Great*	1587
Marlowe, Christopher	*2 Tamburlaine the Great*	1587
Marlowe, Christopher	*Edward the Second*	1592
Marlowe, Christopher; others (?)	*The Jew of Malta*	1589
Middleton, Thomas	*A Chaste Maid in Cheapside*	1613
Middleton, Thomas	*A Mad World, My Masters*	1605
Middleton, Thomas	*A Trick to Catch the Old One*	1605
Middleton, Thomas	*Your Five Gallants*	1607
Shakespeare, William	*The Comedy of Errors*	1594
Shakespeare, William	*Richard the Third*	1592
Shakespeare, William	*The Taming of the Shrew*	1591
Shakespeare, William	*The Two Gentlemen of Verona*	1590

As noted in the Introduction, PCA has been widely adopted as a method for stylistic investigation.[10] Its use in authorship attribution relies on the fact that, when analysing word-frequency counts across a mixed corpus of texts known to be of different authorship, the strongest factor that emerges in the relationship between the texts is generally authorial in nature. Other stylistic signals may also be present, such as the effect of genre, period of composition, gender of the author, and so on, but these are usually demonstrably weaker. For example, Table 1.1 lists a selection of plays by John Lyly, Christopher Marlowe, Thomas Middleton, and William Shakespeare, representing a range of genres and dates of first performance.[11]

With this corpus of machine-readable texts, prepared as outlined above, we use Intelligent Archive, a software tool developed by the Centre for Literary and Linguistic Computing at the University of Newcastle, to generate word-frequency counts for the 500 most frequent words across the corpus, segmented into 2,000-word non-overlapping blocks and discarding any smaller blocks that remain. Proper nouns, foreign-language

[10] See José Nilo G. Binongo and M. W. A. Smith, 'The Application of Principal Component Analysis to Stylometry', *Literary and Linguistic Computing* 14.4 (1999), 445–65.

[11] Appendix A provides further bibliographical details for these plays, including the source texts used and date of publication. The text of *The Jew of Malta* we used excludes the prologues and epilogues attributed to Thomas Heywood. On the possibility of further non-Marlovian revision, see D. J. Lake, 'Three Seventeenth-Century Revisions: *Thomas of Woodstock*, *The Jew of Malta*, and *Faustus B*', *Notes & Queries* 30.2 (1983), 133–43.

https://doi.org/10.1017/9781108120456.002 Published online by Cambridge University Press

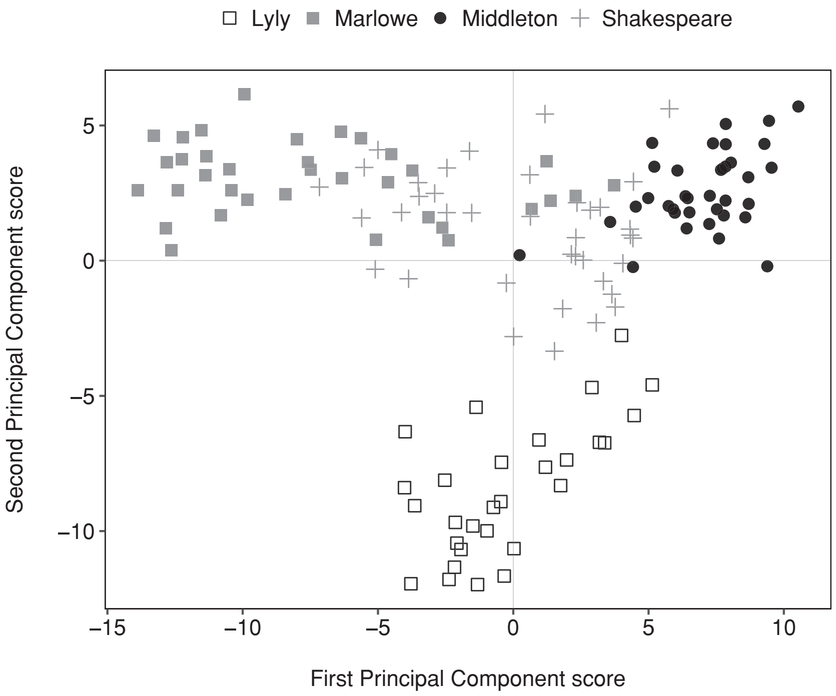

Figure 1.1 PCA scatterplot of 2,000-word non-overlapping segments of plays listed in Table 1.1, using the 500 most frequent words.

words, and stage directions are excluded from the procedure. The result is a large table, with 138 rows (one for each 2,000-word block) and 500 columns (one each for the total of each block's occurrences of each word counted). As one might expect, words such as *the*, *and*, *I*, *to*, and *a* – that is, function words – are among the most frequent.

If it were possible to visualise, and effectively comprehend, every 2,000-word segment could be plotted as a point on a graph along 500 separate axes or dimensions in space. With PCA, we can reduce the dimensionality of the data while preserving as much of the variance as possible. If we use PCA to reduce the data to the two strongest factors, we can then project each 2,000-word segment into a two-dimensional space as a data-point, treating the scores for each segment on the first and second principal components as Cartesian coordinates (Figure 1.1).[12]

[12] Coordinates are a set of numbers that define position in space relative to an origin. In the Cartesian coordinate system, the origin is a fixed point from which two or more axes or 'dimensions' are

https://doi.org/10.1017/9781108120456.002 Published online by Cambridge University Press

The first principal component (the *x*-axis) is the most important latent factor in the various correlations between the word-variables in the segments, and the second principal component (the *y*-axis) is the second most important (independent) latent factor. The relative distances between the points or 'observations' within this space represent degrees of affinity, so that segments of similar stylistic traits – specifically, similar rates of occurrence of our 500 words – cluster tightly together, whereas dissimilar segments are plotted further apart.

To make it easier to read the scatterplot, we use different symbols to label segments belonging to different authors. Although the separation between them is not perfect, segments belonging to the same author tend to cluster together, with Marlowe's segments (plotted as grey squares) typically scoring low on the first principal component and high on the second principal component. The PCA algorithm determines a weighting for each word, negative or positive, to give the best single combination to express the collective variability of all 138 segments' word uses. Marlowe's low score for the first principal component means that the Marlowe segments relatively rarely use the words with a high positive weighting on this component and relatively frequently use the words with a high negative weighting. The algorithm then identifies a second set of weightings for the words, to best account for the remaining collective variability of the 138 segments' word uses after the first principal component has accounted for its fraction of the collective variability. The Marlowe segments use the words with high positive weightings on this second principal component relatively often and the words with high negative ratings relatively rarely. We could, in theory, go on to calculate further principal components (a third, a fourth, and so on) until we run out of variance in the data – which must in any case happen for this experiment when we calculate the 500th principal component and so exhaust our 500 variables' capacity to differ from one another. The most important consideration here is that this method demonstrably captures the affinity of segments by single authors, with Middleton's segments (plotted as black circles) typically scoring high on both axes. Lyly's segments cluster away from the others, scoring comparatively low on the second principal component, whereas Shakespeare's segments gravitate towards the centre of the scatterplot, forming a stylistic 'bridge' between the other authors.

We have plotted the first and second principal components – those which account for the greatest and second greatest proportion of the

defined, with each axis perpendicular to the other. Readers may recall charting plots on graph paper in school mathematics in this same way. We use a Cartesian coordinate system throughout this book to generate scatterplots along two axes – the horizontal or *x* axis, and the vertical or *y* axis.

https://doi.org/10.1017/9781108120456.002 Published online by Cambridge University Press

variance – and what emerge there are separations by author. This is evidence that authorship is a more important factor in stylistic differentiation than other groupings, such as genre or date, as we show below. However, there may be times when we may not be interested in the most important factors, whatever they may be. Since PCA can create as many components as there are variables, it is possible to target a particular factor. If we were interested in date, for example, we could work through the other components to find one which differentiates the sample by date – that is, which single set of weightings given to the 500 words will separate those favoured early in the period from those favoured late in the period – and then either use that component to classify a sample of unknown date, or explore the stylistics of the date-based groupings by examining the patterns of word-variables that create the component.

A different pattern in the data of the first two principal components emerges if we simply re-label the points on the scatterplot according to genre (Figure 1.2). The underlying data has not changed, only the labels of the points. Along the first principal component, segments appear to cluster in generic groups from 'heroical romance' (plotted as black circles) through to 'history' (grey plus symbols), 'tragedy' (unfilled triangles), and 'comedy' (unfilled squares). This perhaps explains some of the internal variation evident within the authorial clusters identified in Figure 1.1. For example, segments from Marlowe's 'heroical romance' plays, *1* and *2 Tamburlaine the Great*, cluster tightly together, whereas segments from his *Edward the Second* are plotted closer to – sharing stylistic traits with – segments from Shakespeare's play of the same genre, *Richard the Third*. Similarly, Lyly's 'comedy' *Mother Bombie* is plotted higher on the second principal component than segments from his other 'classical legend' comedies.[13]

PCA works by finding weightings for the variables to establish new composite variables – the components. We can examine these weightings to find out which variables contribute the most to a given component. To visualise the weightings, we can plot them in a separate biaxial chart, show them as a column or bar chart, or display them on the same chart as the segments in the form of a 'biplot'. The biplot allows us to visualise the contributions of each word-variable in the same two-dimensional space as the play segments (Figure 1.3).[14] It shows the segment scores (as in Figures 1.1

[13] We also re-labelled the data-points by decade of first production. There were some clusters, but a far less clear-cut division than by author or genre. Plays of the 1590s occupied the middle part of the first principal component, but 1580s plays overlapped them substantially, and 1610s plays were all within the range of the 1600s plays. Plays of the 1580s were spread over almost the full range of the second principal component.

[14] Michael Greenacre, *Biplots in Practice* (Bilbao: Fundación BBVA, 2010), 15–24, 59–68; Alt, *Exploring Hyperspace*, 92–7.

https://doi.org/10.1017/9781108120456.002 Published online by Cambridge University Press

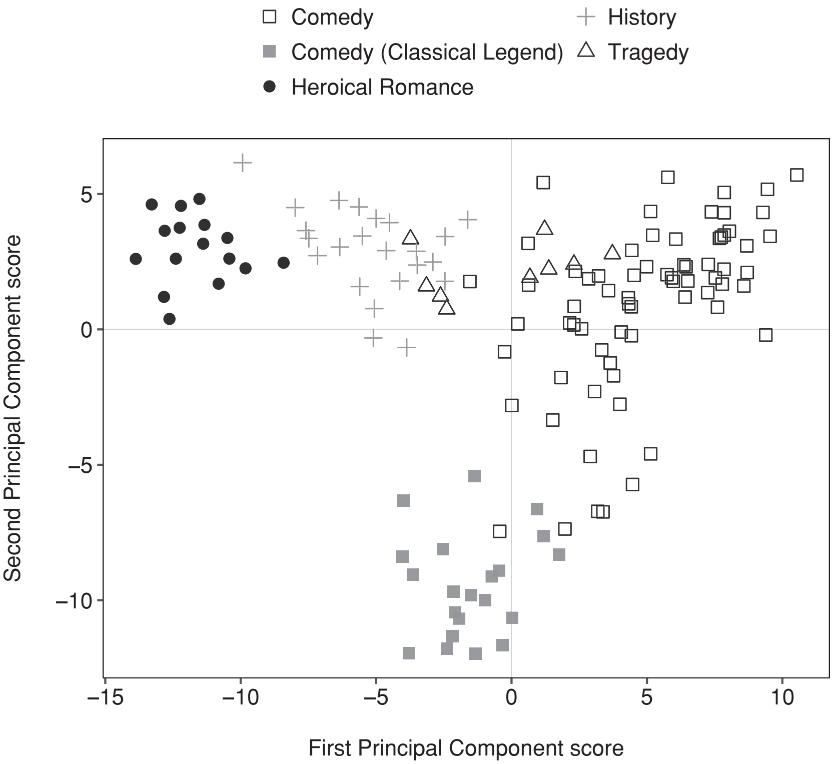

Figure 1.2 PCA scatterplot of 2,000-word non-overlapping segments of plays listed in Table 1.1, labelled by genre, using the 500 most frequent words.

and 1.2), but also overlays on these the weightings for the word-variables. In a biplot, the word-variables are generally represented by an arrow or 'vector' drawn from the origin – the point where the *x* and *y* axes intersect, i.e., 0,0 – rather than as points. This is a reminder that each variable is an axis, and the length and direction of the vector is also a convenient indication of the importance of that variable for a given component.

The positions of the ends of the vectors in the biplot are determined by the weightings of the variables for the two components, re-scaled to fit into the chart space.[15] Since the vectors are scaled to fit the biplot, the distance between the end or 'head' of a vector and a play segment is unimportant; what matter are the directions and relative lengths of the vectors.

[15] The biplots in this book were produced with the R statistical computing package, using the default scaling factors.

https://doi.org/10.1017/9781108120456.002 Published online by Cambridge University Press

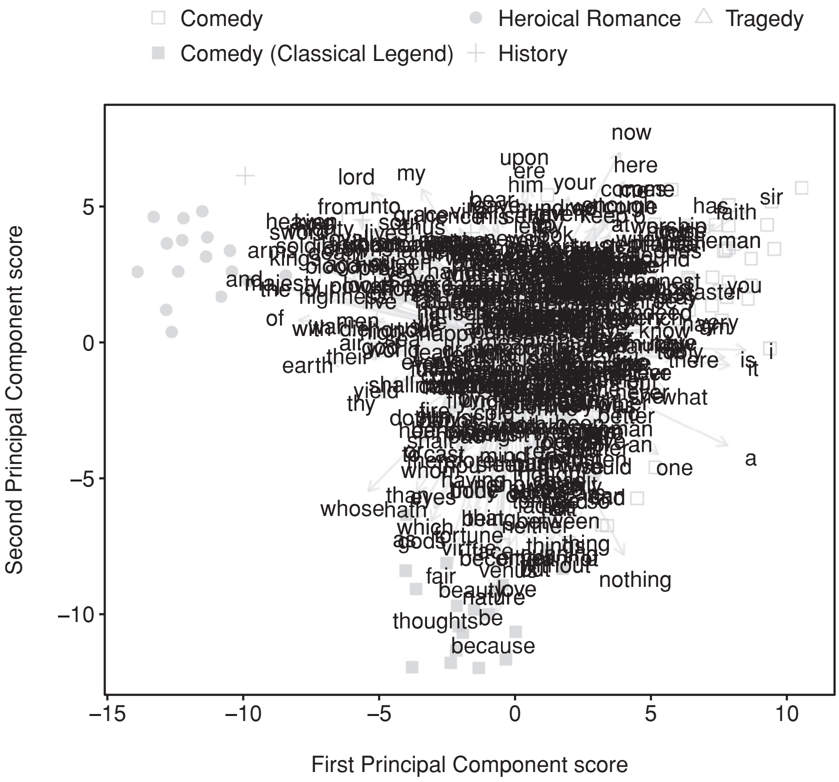

Figure 1.3 PCA biplot of 2,000-word non-overlapping segments of plays listed in Table 1.1, labelled by genre, using the 500 most frequent words.

The direction of a vector indicates how a word-variable contributes to each of the principal components. A segment with many instances of the word-variables strongly positively weighted in one of the principal components will have been 'driven' in that direction, whereas a segment dominated by word-variables weak in both components will be plotted towards the origin.[16] The relative length of a vector corresponds to the magnitude of the contribution. In Figure 1.3, a long vector extending in an easterly direction shows that the corresponding word-variable has a heavy positive

[16] To increase legibility, PCA biplots often omit the vectors and plot only the word-variable labels, projected as points. The result is the same: the word-variables are plotted by their weightings on the two principal components, so that word-variables appearing to the extremes of the axes are those that make the most difference in the scatter of the segments along the axes. In our book, we have sometimes created a separate chart of the variables or a selection of variables for increased legibility.

https://doi.org/10.1017/9781108120456.002 Published online by Cambridge University Press

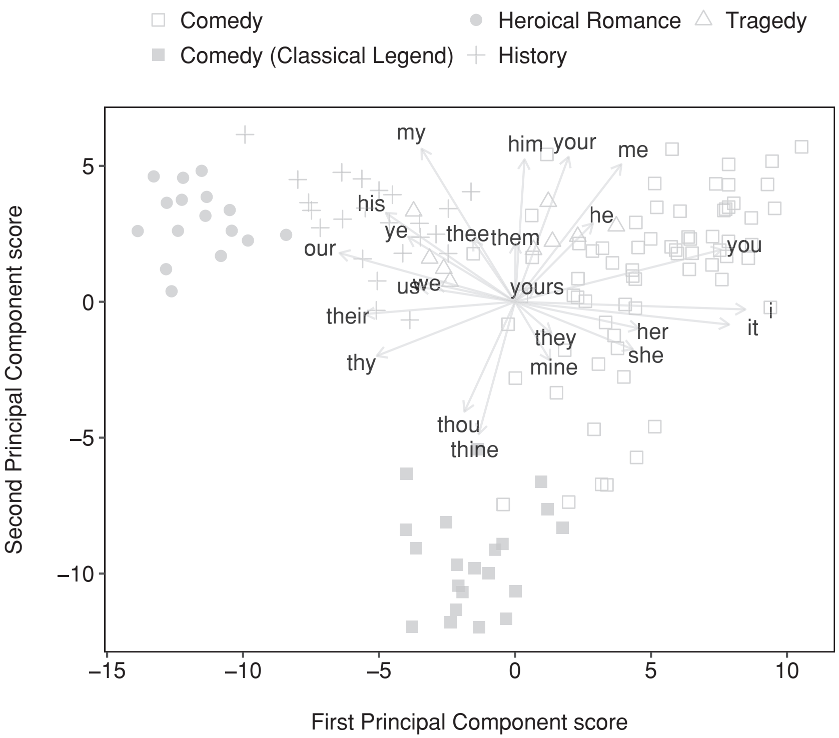

Figure 1.4 PCA biplot of 2,000-word non-overlapping segments of plays listed in Table 1.1, labelled by genre, using the 500 most frequent words and highlighting personal pronouns.

weighting on the first principal component, while a short vector extending in a southerly direction shows that the corresponding word-variable has a weak negative weighting on the second principal component.

If, as in Figure 1.3, all 500 of the word-variable vectors are drawn, the biplot becomes too difficult to analyse. Instead, we can redraw the biplot highlighting only word-variables of thematic interest or those belonging to a particular grammatical class. For example, Figure 1.4 gives the same biplot with only vectors for word-variables of personal pronouns drawn.

Inspection of the biplot reveals that 'comedy' segments plotted to the east of the origin are dominated by singular personal pronouns, such as the first-person *I*, *me*, and *mine*, the second-person formal *you*, *your*, and *yours*, and the third-person *he*, *she*, *it*, *him*, and *her*. By contrast, the 'heroical romance' and 'history' segments plotted west of the origin are dominated by plural personal nouns, such as the first-person *we*, *us*, and *our*, the second-person

https://doi.org/10.1017/9781108120456.002 Published online by Cambridge University Press

ye, and the third-person *their*, while 'classical legend' comedy segments, plotted south-west of the origin, favour the second-person informal singular *thou* and *thine* forms. Speeches in heroical romances, tragedies, and history plays are evidently cast more in terms of collectives, as we might expect with a focus on armies in battle and political factions. Comedies, on the other hand, tend to include more one-on-one interpersonal exchanges, so that the singular pronouns figure more strongly in their dialogue.

Random Forests

The decision-making process is often characterised as a series of questions: answers to one question may lead to a decision being reached, or prompt a further question – or series of questions – until a decision is made. For example, a doctor asks a patient to describe their symptoms and they respond that they have a runny nose. Among other conditions, rhinorrhea – the technical term for a runny nose – is a symptom common to both allergy (e.g. hayfever) and certain infections (e.g. the common cold). To reach a diagnosis, the doctor may ask further questions of the patient: how long have the symptoms persisted? Is the nasal discharge clear or coloured? Does the patient suffer from itchy eyes, aches, or fever?

While the common cold often causes a runny nose and may sometimes occasion aches, it rarely results in fever or itchy eyes and typically does not last longer than a fortnight. By contrast, rhinorrhea and itchy eyes are frequent allergic reactions and may last as long as the patient is in contact with the allergy trigger – minutes, hours, days, weeks, even months and seasons. (The term 'hayfever' is somewhat misleading, because fever and aches are not typical allergic responses.) Our hypothetical doctor's decision-making process may be visualised as a decision tree (Figure 1.5).

Of course, this is a simplified example (and correspondingly simple visualisation) of a complex consideration of multiple variables, some of which are 'weighted' – or more important to the decision-making process than others.

Random Forests is a supervised machine-learning procedure for classifying data using a large number of decision trees.[17] Whereas our hypothetical doctor relied upon centuries of accumulated knowledge to identify

[17] Leo Breiman, 'Random Forests', *Machine Learning* 45.1 (2001), 5–32. Much of the description that follows appeared in an earlier form as part of Jack Elliott and Brett Greatley-Hirsch, '*Arden of Faversham*, Shakespeare, and "the print of many"', in Gary Taylor and Gabriel Egan (eds.), *The New Oxford Shakespeare: Authorship Companion* (Oxford University Press, 2017), 139–81. We thank the editors for their permission to reproduce and adapt those passages here.

https://doi.org/10.1017/9781108120456.002 Published online by Cambridge University Press

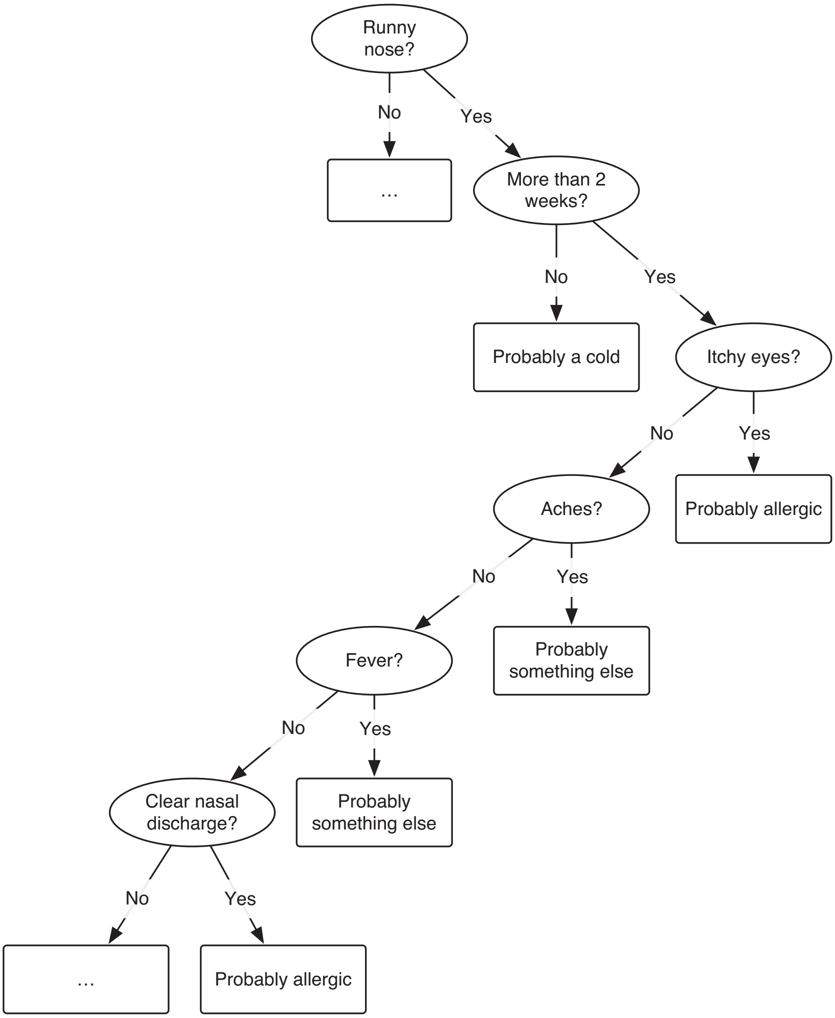

Figure 1.5 Binary decision tree diagram.

attributes or features distinguishing one medical condition from another, decision tree algorithms instead begin by testing variables in a set of data with a known shared attribute (a so-called 'training set') to derive a rule – like the series of questions posed by the doctor – that best performs the task of splitting the data into desired categories or classes. At each succeeding level of the tree, the sub-sets created by the splits are themselves split

https://doi.org/10.1017/9781108120456.002 Published online by Cambridge University Press

according to another rule, and the tree continues to grow in this fashion until all of the data has been classified. Once a decision tree is 'trained' to classify the data of the training set, it can then be employed to classify new, unseen data.[18]

Random Forests combines the predictive power of hundreds of such decision trees (hence 'forests'). Each tree is derived using a different and random sub-set of the training dataset and variables. To enable validation of the technique and to avoid the problem of 'over-fitting',[19] randomly selected segments of the training set are withheld from the algorithm so that they do not inform the construction of the decision trees (and thus allowing us to determine how accurate the trees' predictions are in classifying these withheld segments). By default, one-third of all training-set segments are withheld for this purpose. This testing, using segments of a known class or category, treated as if this was unknown, gives us an expected error rate for when the decision trees are used to classify new data. The higher the classification error rate, the weaker the relationship between the variables and the classes, and *vice versa*. Hundreds of such trees are constructed, and for each classification to be made each tree contributes one vote to the outcome. This aggregation of decision trees evens out any errors made by individual trees that may arise from the construction of apparently reliable – but in fact false – rules based on anomalous data.[20]

By way of example, we use Intelligent Archive to generate word-frequency counts for the 500 most frequent words across the selection of plays listed in Table 1.1, segmented into 2,000-word non-overlapping blocks and discarding any smaller blocks that remain. As before, proper nouns, foreign-language passages, and stage directions are excluded from the procedure. This produces a large table of 138 rows and 500 columns, which we split into two separate tables: one to serve as our training dataset

[18] As such, the procedure is 'supervised' because the algorithm relies upon human pre-processing of the training-set data to ensure that it is characterised by a shared attribute, such as particular medical conditions, or, for our purposes, play-texts of common authorship, genre, period, or repertory company.

[19] 'Over-fitting' occurs when a machine-learning algorithm or statistical model performs well on the training data, but generalises poorly to any new data. To classify training data on a two-dimensional chart, for example, we may use a highly complex equation to generate a wavy line snaking around each data-point to serve as a boundary between groups. While this equation might perfectly separate the data-points into groups, it may also 'fit' or reflect the exact contours of the training data too closely. A simpler equation, producing a line with a looser fit to the training data, may better serve as a boundary when we wish to classify newly introduced data.

[20] For example, a decision tree derived from analysis of a patient suffering from a runny nose caused by an unusually resilient and long-lasting cold might generate the rule 'If symptoms persist for longer than two weeks, then it is a cold'. While accurate in the case of this particular, local anomaly, this rule does not perform well as a predictor for the majority of cases.

https://doi.org/10.1017/9781108120456.002 Published online by Cambridge University Press

(109 rows, 500 columns), and another containing all of the segments from each author's first-listed play to serve as a test dataset (29 rows, 500 columns). After a randomly selected one-third of segments in the training dataset are withheld by the algorithm to be tested later, 500 decision trees are populated using the remaining two-thirds of the training dataset, trying 22 random word-variables at each 'split' in the decision tree.[21] A diagram of one of the decision trees populated in this experiment is given in Figure 1.6, in which the rules are expressed as the rates of occurrence of a word-variable per 2,000 words. Thus, according to its rules, if a 2,000-word segment contains 1 or fewer instances of the word *hath* and 61 or fewer instances of the word *and*, then this decision tree predicts it is a Middleton segment.[22] Of course, as the outcome of a single decision tree, this prediction would count as one out of 500 votes cast by the 'forest' of trees.

The algorithm then uses the decision trees to classify the training dataset as a whole, with the randomly withheld one-third of segments reintroduced. This produces an expected error rate for when the unseen test dataset will be classified later. Table 1.2 gives the confusion matrix for the 109 segments of the entire training dataset, tabling four misclassifications made by the decision trees: three segments of *The Jew of Malta* assigned to Shakespeare, and one segment of *Richard the Third* assigned to Marlowe. This produces a promisingly low expected error rate of 3.67 per cent (= 4 ÷ 109 × 100).

The decision trees are then used to classify all of the data – i.e., the whole training dataset, including the previously withheld segments, as well as the newly introduced segments of the test dataset. Table 1.3 gives the resulting confusion matrix. The decision trees classify all of the segments in the test dataset correctly, resulting in a classification error rate of 2.89 per cent for all of the segments in both the training and test datasets – that is, 4 misclassified segments out of the total 138.

[21] A function built into the Random Forests algorithm compares estimated error rates when different values for the number of variables are tried at each split and selects the optimal value (i.e., the value resulting in the lowest expected error rate). By default, the first number of variables tried is the square root of the total number of variables, rounded down to the nearest whole number – in our case 22 (the approximate square root of 500). The algorithm then generates other values to try by multiplying or dividing the first number by a factor – by default, this factor is 2. New values are continuously tried so long as the expected error rate improves beyond a given threshold (by default, 5 per cent). Here and elsewhere in this book, we use the default settings of the Random Forests algorithm. Thus, in this example, the algorithm first tries 22 variables at each split, and then compares the estimated error rates when 6 (or 22 ÷ 2 ÷ 2), 11 (or 22 ÷ 2), 44 (or 22 × 2), 88 (22 × 2 × 2), and 176 (or 22 × 2 × 2 × 2) variables are tried. Of these, 22 is determined the optimal value.

[22] Although the actual split for *hath*, as per the diagram, is a rate of ≥ 1.5 instances per 2,000 words, our word-frequency counts are given only in whole, discrete numbers. Since we cannot have 1.5 instances of *hath*, in practice the rule applies to ≥ 1 instances in a 2,000-word segment.

https://doi.org/10.1017/9781108120456.002 Published online by Cambridge University Press

https://doi.org/10.1017/9781108120456.002 Published online by Cambridge University Press

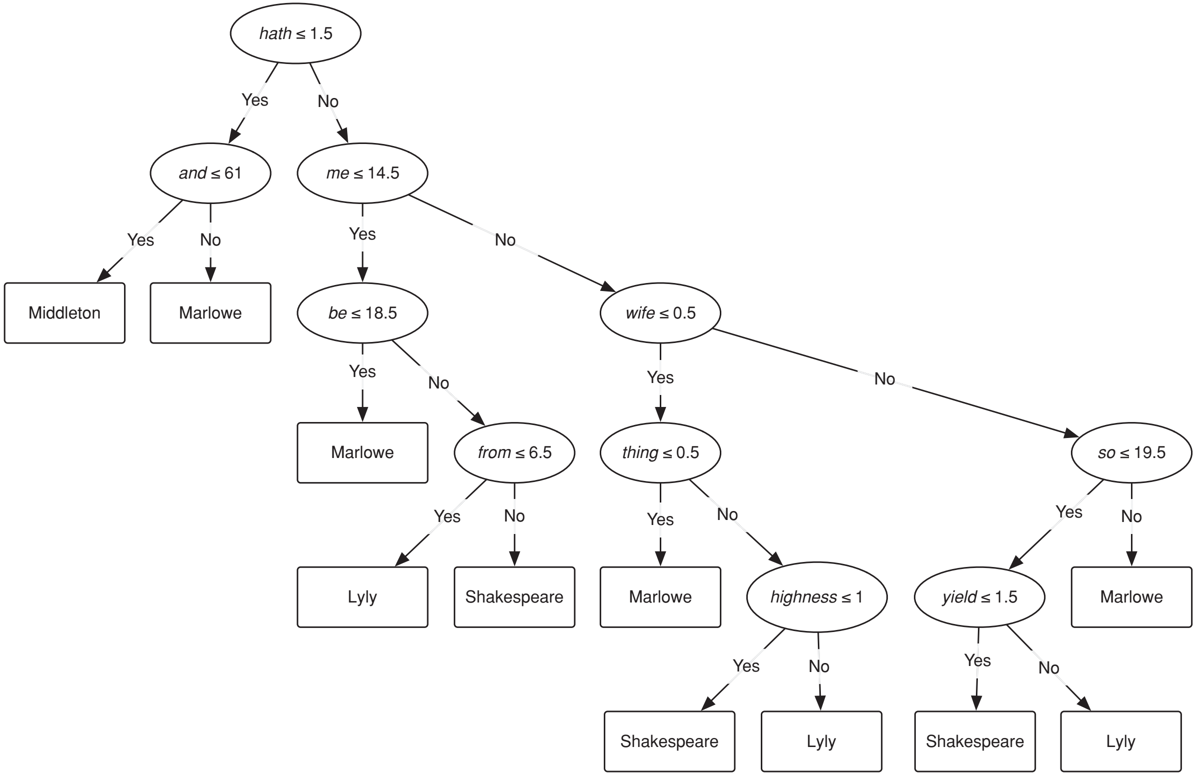

Figure 1.6 Diagram of a single binary decision tree populated for Random Forests classification of 2,000-word non-overlapping segments in a training dataset of 109 segments drawn from plays listed in Table 1.1, using the 500 most frequent words.

Table 1.2 *Confusion matrix for Random Forests classification of 2,000-word non-overlapping segments in a training dataset of 109 segments drawn from plays listed in Table 1.1, using the 500 most frequent words*

	Lyly, John	Marlowe, Christopher	Middleton, Thomas	Shakespeare, William	Misclassification (%)
Lyly, John	23	0	0	0	0
Marlowe, Christopher	0	24	0	3	11
Middleton, Thomas	0	0	28	0	0
Shakespeare, William	0	1	0	30	3

Delta

Delta is a supervised method introduced by John Burrows to establish the stylistic difference between two or more texts by comparing the relative frequencies of very common words.[23] Although well established as a tool for authorship attribution study, Delta is also used more broadly as a means to describe 'the relation between a text and other texts in the context of the entire group of texts'.[24]

In its usual deployment, the procedure establishes a series of distances between a single text of interest and a comparison set typically comprising a series of authorial sub-sets of texts. The author with the lowest distance score is judged to be the 'least unlikely' author of the mystery text.[25] There are two main steps. The procedure begins by generating counts of

[23] John Burrows, 'Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship', *Literary and Linguistic Computing* 17.3 (2002), 267–86, and 'Questions of Authorship: Attribution and Beyond', *Computers and the Humanities* 37.1 (2003), 5–32. For assessments of the method, see David L. Hoover, 'Testing Burrows's Delta', *Literary and Linguistic Computing* 19.4 (2004), 453–75, and Shlomo Argamon, 'Interpreting Burrows's Delta: Geometric and Probabilistic Foundations', *Literary and Linguistic Computing* 23.2 (2008), 131–47. A number of refinements of Delta have been proposed for the purpose of authorship attribution; see, for example, Peter W. H. Smith and W. Aldridge, 'Improving Authorship Attribution: Optimizing Burrows' Delta Method', *Journal of Quantitative Linguistics* 18.1 (2011), 63–88. However, for simplicity, we here describe the original version as proposed by Burrows.

[24] Fotis Jannidis and Gerhard Lauer, 'Burrows's Delta and Its Use in German Literary History', in Matt Erlin and Lynne Tatlock (eds.), *Distant Readings* (Rochester: Camden House, 2014), 32.

[25] Sections of the description that follows appeared in an earlier form as part of Jack Elliott and Greatley-Hirsch, '*Arden of Faversham*'. We thank the editors for their permission to reproduce and adapt those passages here.

https://doi.org/10.1017/9781108120456.002 Published online by Cambridge University Press

Table 1.3 *Confusion matrix for Random Forests classification of 109 training and 29 test segments of plays listed in Table 1.1, segmented into 2,000-word non-overlapping blocks, using the 500 most frequent words*

	Lyly, John	Marlowe, Christopher	Middleton, Thomas	Shakespeare, William	Misclassification (%)
Lyly, John	29	0	0	0	0
Marlowe, Christopher	0	32	0	3	8
Middleton, Thomas	0	0	36	0	0
Shakespeare, William	0	1	0	37	2

high-frequency words in the 'test' text and comparison set. Counts for individual texts in the comparison set are retained, allowing Delta to derive both a mean figure for the set as a whole, and a standard deviation – a measure of the variation from that mean – for each variable.[26] The counts on the chosen variables – usually very common words – are transformed into percentages to account for differing sizes of text and then into *z*-scores by taking the difference between the word counts and the mean of the overall set and dividing that by the standard deviation for the variable. Using *z*-scores has the advantage that low-scoring variables are given equal weight with high-scoring ones, since a *z*-score is the number of standard deviations of an observation from the mean, unrelated to the size of the original units. The *z*-score also takes into account the amplitude of fluctuations within the counts. Wide fluctuations result in a high standard deviation and thus a lower *z*-score.

The differences between *z*-scores for the test text and each authorial subset are then found for each variable, adding up the absolute differences – that is, ignoring whether the figures are positive or negative – to form a

[26] The four-figure sets $\{6, 7, 2, 9\}$ and $\{8, 1, 3, 12\}$ both have a mean of 6, since this is one-fourth of $6 + 7 + 2 + 9 \ (= 24)$ also one-fourth of $8 + 1 + 3 + 12 \ (= 24)$. However, the figures in the second set differ more widely from their mean than those in the first set. To express this, the standard deviation for each set is derived by squaring each data-point's difference from its set's mean, dividing the resulting squares by $(N - 1)$, i.e., the number of samples less one, and then finding the square root of that number. For the first set, this is the square root of one-third of $(6 - 6)^2 + (7 - 6)^2 + (2 - 6)^2 + (9 - 6)^2$, which comes to about 2.9. For the second set, this is the square root of one-third of $(8 - 6)^2 + (1 - 6)^2 + (3 - 6)^2 + (12 - 6)^2$, or roughly 5. These are 'sample standard deviations' – i.e., the standard deviations of samples understood to be representing larger populations. This is the version of the metric we use in the studies in this book.

https://doi.org/10.1017/9781108120456.002 Published online by Cambridge University Press

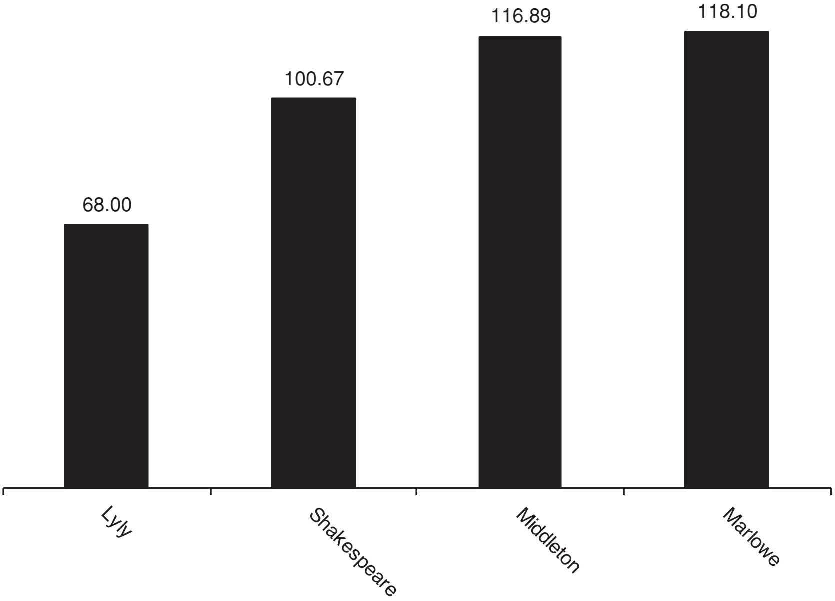

Figure 1.7 Delta distances between *Galatea* and four authorial sub-sets.

composite measure of difference (or 'Delta' distance). The procedure is complete at this point, with a measure for the overall difference between the test text and each of the authorial sub-sets within the comparison set.

To illustrate the method, consider an example using the set of sixteen plays listed in Table 1.1, with four plays each by Lyly, Marlowe, Middleton, and Shakespeare. We first generate frequency counts for the 100 most common function words in all 16 plays and transform these into percentages. We then choose one Lyly play at random to serve as a test text – in this case, *Galatea* – and withdraw this play from the Lyly authorial sub-set. We transform the word-frequency scores for *Galatea* into *z*-scores, using the means and standard deviations for the whole set of sixteen plays. We do the same for the mean scores for the Lyly, Marlowe, Middleton, and Shakespeare plays – the Lyly set consisting of the three remaining plays, the others retaining their full sub-set of four plays each. To arrive at a composite distance measure, we add up the absolute differences between the *Galatea* *z*-scores and each of the authorial sub-set *z*-scores for the 100 word-variables. Figure 1.7 shows the resulting Delta distances as a column chart.

https://doi.org/10.1017/9781108120456.002 Published online by Cambridge University Press

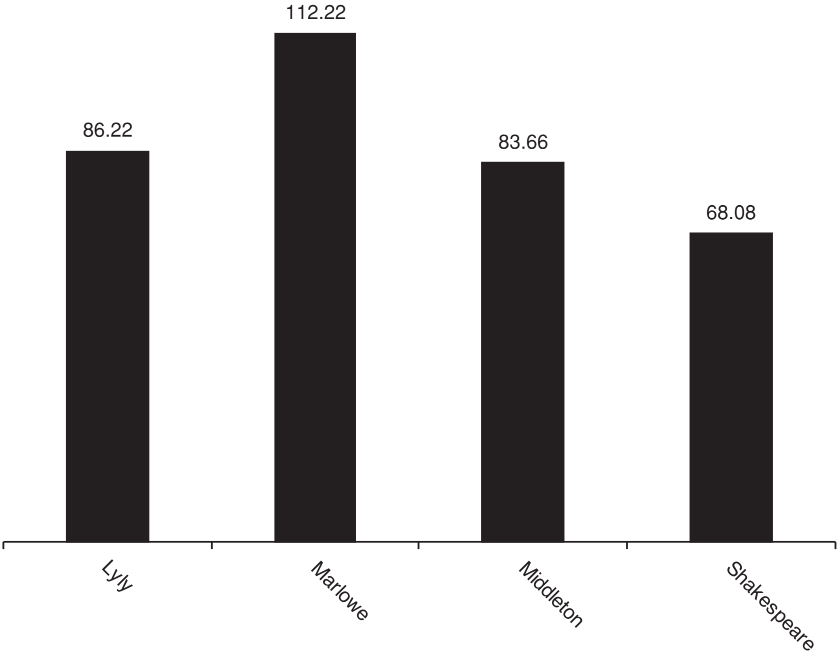

Figure 1.8 Delta distances for *The Jew of Malta* and four authorial sub-sets.

When treated as a mystery text, *Galatea* finds its closest match in a Lyly sub-set based on the three remaining Lyly plays, with a Delta distance of 68. Shakespeare, with a Delta distance of 100.67, is the next nearest author, followed by Middleton (116.89) and Marlowe (118.10). We can then do the same for the other three Lyly plays, withdrawing each in turn and testing the resemblance between that play and each of the four authorial sub-sets. As it turns out, and as we would expect (but could not guarantee), each Lyly play matched the sub-set of remaining Lyly plays most closely.

We repeat the procedure for the other authors along the same lines, holding out and testing each play in turn. In every case, the known author was the closest match, with the exception of *The Jew of Malta* (Figure 1.8), which matched Shakespeare most closely (with a Delta distance of 68.08), then Middleton (at 83.65), then Lyly (at 86.22) – with Marlowe the most distant at 112.22.[27]

[27] It is worth noting that the Random Forests algorithm, outlined above, also classified segments of *The Jew of Malta* as Shakespeare's.

https://doi.org/10.1017/9781108120456.002 Published online by Cambridge University Press

Evidently, this play represents a radical departure from Marlowe's typical practice in the use of very common function words (established on the basis of the three other plays in the sub-set). As the only (potentially) incorrect attribution out of sixteen Delta tests, this anomalous result is certainly worthy of further investigation. However, for the purposes of demonstration, it is enough to note that, overall, Delta is a good – but perhaps not infallible – guide to authorship and stylistic difference, even when using small sub-sets to represent an author.

Shannon Entropy

Shannon entropy is a measure of the repetitiveness of a set of data, and is the key concept in information theory as developed by Claude Shannon in the 1940s.[28] Shannon entropy calculates the greatest possible compression of the information provided by a set of items considered as members of distinct classes. A large entropy value indicates that the items fall into a large number of classes, and thus must be represented by listing the counts of a large number of these classes. In an ecosystem, this would correspond to the presence of a large number of species each with relatively few members. The maximum entropy value occurs where each item represents a distinct class. Minimum entropy occurs where all items belong to a single class. In terms of language, word tokens are the items and word types the classes.[29] A high-entropy text contains a large number of word types, many with a single token. A good example would be a technical manual for a complex machine which specifies numerous distinct small parts. A low-entropy text contains few word types, each with many occurrences, such as a legal document where terms are repeated in each clause to avoid ambiguity. Entropy is a measure of a sparse and diverse distribution versus a dense and concentrated one. High-entropy texts are demanding of the reader and dense in information – they constantly move to new mental territories; they are taxing and impressive. Low-entropy texts are reassuring and familiar – they are implicit in their signification, assuming common

[28] C. E. Shannon, 'A Mathematical Theory of Communication', *Bell System Technical Journal* 27 (1948), 379–423, and 'Prediction and Entropy of Printed English', *Bell System Technical Journal* 30 (1951), 50–64. For a more accessible overview of entropy and information theory, see Luciano Floridi, *Information: A Very Short Introduction* (Oxford University Press, 2010), 37–47; and James Gleick, *The Information: A History, a Theory, a Flood* (New York: Pantheon, 2011), 204–32.

[29] On the application of Shannon entropy and other measures to literary study, see Osvaldo A. Rosso, Hugh Craig, and Pablo Moscato, 'Shakespeare and Other English Renaissance Authors as Characterized by Information Theory Complexity Quantifiers', *Physica A* 388 (2009), 916–26.

https://doi.org/10.1017/9781108120456.002 Published online by Cambridge University Press

knowledge, while high-entropy texts specify and create contexts for themselves. High-entropy texts contain more description and narrative, while low-entropy texts contain more dialogue.

Shannon entropy is defined as the negative of the sum of the proportional counts of the variables in a dataset each multiplied by its logarithm.[30] A line consisting of a single word-type repeated five times (e.g. 'Never, never, never, never, never!' *King Lear* 5.3.307) has a single variable with a proportion of $\frac{5}{5}$ (or 1). The log of 1 is 0. The Shannon entropy of the line is therefore:

$$-\left(\frac{5}{5}\log\frac{5}{5}\right) = 0$$

The line 'Tomorrow, and tomorrow, and tomorrow' from *Macbeth* (5.5.18) has three instances of *tomorrow* and two of *and*. The proportional count for *tomorrow* is $\frac{3}{5}$ (or 0.6) and for *and* is $\frac{2}{5}$ (or 0.4), thus the Shannon entropy for the line is

$$-\left[\left(\frac{3}{5}\log\frac{3}{5}\right) + \left(\frac{2}{5}\log\frac{2}{5}\right)\right] \approx 0.673$$

For a final comparison, consider the line: 'If music be the food of love, play on' (*Twelfth Night* 1.1.1). This time, each of the nine words making up the line occurs only once. Since each word-variable has a proportional score of $\frac{1}{9}$ (or $\approx$ 0.111), the Shannon entropy for this line is:

$$-\begin{bmatrix} \left(\frac{1}{9}\log\frac{1}{9}\right) + \left(\frac{1}{9}\log\frac{1}{9}\right) + \left(\frac{1}{9}\log\frac{1}{9}\right) + \left(\frac{1}{9}\log\frac{1}{9}\right) + \left(\frac{1}{9}\log\frac{1}{9}\right) \\ + \left(\frac{1}{9}\log\frac{1}{9}\right) + \left(\frac{1}{9}\log\frac{1}{9}\right) + \left(\frac{1}{9}\log\frac{1}{9}\right) + \left(\frac{1}{9}\log\frac{1}{9}\right) \end{bmatrix} \approx 2.197$$

– a higher score than for our two previous examples, reflecting comparatively greater variability in word use.[31]

[30] The formula to derive the Shannon entropy (H) for X is:

$$H(X) = -\sum_{x_i \in X} x_i \log x_i$$

A logarithm represents the power to which a fixed number or base must be raised to produce a given number. In all of our Shannon entropy calculations, we use natural logarithms, where the base is *e*, approximately 2.718. Because the logarithm of a fraction (as all proportions are) is always negative, the Shannon entropy formula calls for the negative of the sum ($-\Sigma$) of these proportional counts multiplied by their logarithms ($x_i \log x_i$) to ensure that the result is positive.

[31] Shannon entropy is sensitive to text length – the maximum possible entropy increases as text length increases. To account for this, we work with samples of the same length when we go beyond the illustrative examples given here.

https://doi.org/10.1017/9781108120456.002 Published online by Cambridge University Press

t-tests

Consider the following experiment. We compile a set of Shakespeare's comedies (*All's Well That Ends Well*, *As You Like It*, *The Merchant of Venice*, *A Midsummer Night's Dream*, *Much Ado About Nothing*, and *Twelfth Night*) and a set of Shakespeare's tragedies (*Antony and Cleopatra*, *Hamlet*, *King Lear*, *Othello*, *Romeo and Juliet*, and *Troilus and Cressida*). Are there more instances, on average, of the word *death* in the tragedies compared with the comedies? If there is a difference in these averages, how consistent is it, in the sense that any large group of tragedies will have more occurrences of *death* overall? Our experiment calls for us to find a way to see past mere averages to the varying counts that lie behind them.

The occurrence of *death* in each of these plays, expressed as a percentage of the total number of words, is 0.08, 0.03, 0.05, 0.08, 0.08, and 0.05 respectively for the comedies, and 0.14, 0.13, 0.08, 0.06, 0.29, and 0.06 for the tragedies. The mean for the comedies is 0.062, and for the tragedies it is 0.127 – more than twice as large. But how can we take the fluctuations within the groups into account?

One way to do this is to use a *t*-test, a common statistical procedure to determine whether the 'mean' or average of a 'population' – that is, all members of a defined group or dataset from which a selection or 'sample' is drawn – differs significantly from a hypothetical mean or the mean of another population. The test was first proposed in 1908 by W. S. Gosset, writing under the pseudonym 'Student' while working in quality control for the Guinness brewery in Ireland.[32] Student's *t*-test, as it has come to be known, generates a simple metric called the *t*-value, calculated as the difference in means between two sets divided by the combination of their standard deviations. A high *t*-test score means that the average use in one set is much higher or lower than the use in a second set, and the word overall does not fluctuate much.

Student's *t*-test assumes that the two populations under investigation follow a 'normal distribution' and have an equal variance (i.e., the data in both populations is 'spread' or 'scattered' equally).[33] In 1947 B. L. Welch adapted Student's *t*-test to accommodate populations of unequal variance,[34] and we

[32] 'Student' [= W. S. Gosset], 'The Probable Error of a Mean', *Biometrika* 6.1 (1908), 1–25.

[33] If plotted on a graph, data with a 'normal distribution' would resemble a symmetrical, bell-shaped curve, with the density of the curve centred about its mean. With an equal 'variance', the data in both populations is 'spread' or 'scattered' equally. (Standard deviation is the square root of the variance.)

[34] B. L. Welch, 'The Generalization of "Student's" Problem When Several Different Population Variances Are Involved', *Biometrika* 34.1–2 (1947), 28–35. We use the two-tailed heteroscedastic version.

https://doi.org/10.1017/9781108120456.002 Published online by Cambridge University Press

use this variation in the present experiment and generally throughout this book. For Welch's *t*-test, the one we have used in this book, the formula is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Here $\bar{x}_1$ and $\bar{x}_2$ are the means of the first and second samples, s_1^2 and s_2^2 the squared standard deviations of the first and second samples, and n_1 and n_2 the number of items in each respective sample. For our experiment, we already have the means (as above, 0.062 and 0.127), and the standard deviations are 0.021 and 0.087 for the comedies and tragedies respectively. The sample size is 6 for both sets. Using these figures, the formula produces a *t*-value of –1.778.

The other necessary piece of information is the number of degrees of freedom in the analysis. The more degrees of freedom, the more information the result is based on and the more confident we can be that the result reflects an underlying truth. Degrees of freedom in the *t*-test depend on the number of samples, but with Welch's *t*-test we are allowing for the possibility of different variances for the two groups, and estimating the true number of degrees of freedom requires taking into account the distribution of the data using the Welch-Satterthwaite formula.[35]

The result in this case is 5.6. Given this number, we can find a *t*-test probability by consulting a table or using a *t*-test probability calculator.[36] This *t*-test probability (or '*p*-value') indicates how often a difference like this would come about merely by chance, even when the two sets in fact belong to the same overall population, given the sample size. For this experiment, using a figure of 5.6 for the relevant degrees of freedom results in a *p*-value of 0.129. This is the probability (given that the data is normally distributed) that the two samples come from the same parent population – that the difference is a matter of local variation rather than something underlying and consistent. That is, 13 per cent of the time (one time in seven or

See George W. Snedecor and William G. Cochran, *Statistical Methods*, 8th edn (Ames: Iowa State College Press, 1989), 53–8.

[35] This is a complicated formula, and researchers normally use a statistics package to find the degrees of freedom in a particular case. Here we use *SPSS*. For the background, see Welch, 'The Generalization', and F. E. Satterthwaite, 'An Approximate Distribution of Estimates of Variance Components', *Biometric Bulletin* 2.6 (1946), 110–14. See also Les Kirkup and Bob Frenkel, 'The *t*-distribution and Welch-Satterthwaite Formula', in *An Introduction to Uncertainty and Measurement* (Cambridge University Press, 2006), 162–90.

[36] Here and elsewhere in this book, we use the TTEST function in Microsoft Excel. Figures will vary when using different *t*-test calculators as a result of how values are rounded.

https://doi.org/10.1017/9781108120456.002 Published online by Cambridge University Press

eight) we should expect to see this apparent difference between the comedies and tragedies purely by chance alone, even if comedies and tragedies have no underlying preference for or against using the word *death*. This suggests that although the tragedies in our sample on average have twice the instances of the word, the fluctuations within the sets and the small number of samples mean that we should not base any broad conclusions on this result.

Glancing at the proportional scores for *death* in these texts might have indicated the same thing. There is one aberrant high score, for *Romeo and Juliet*, which accounts for a great deal of the high average for the set of tragedies overall, and there are three comedies at 0.8, which are all higher than the two lowest-scoring tragedies at 0.6. The *t*-test offers a way to treat these fluctuations systematically, a summary statistic which can be carried over from one comparison to another, and a broad indication about the inferences we can safely make about wider populations (such as about Shakespeare comedy and tragedy in general) from the current sample.

PCA, Random Forests, Delta, Shannon entropy, and the *t*-test are all well-established tools that we have found useful in making sense of the abundant, multi-layered data which can be retrieved from literary texts. They take us beyond what we can readily see with the naked eye, as it were – a count that stands out as high or low, or an obvious pattern of association between variables or samples – to larger-scale, more precise summaries that have some in-built protections from bias. PCA is a data reduction method; Random Forests a classification tool; Delta a distance measure; Shannon entropy a density metric; and the *t*-test takes us back to single variables and the question of whether two sets of counts have an underlying difference, or only an apparent one. They are just five of the numerous methods available, and by no means the most complex, but they are all tried and tested and offer a useful range. They come from different eras and were developed for different purposes – only Delta was devised specifically for computational stylistics. All five can be used both to test a hypothesis and to explore data more inductively, as we demonstrate in the chapters that follow.

https://doi.org/10.1017/9781108120456.002 Published online by Cambridge University Press