CHAPTER 4.2

Computational studies

BRETT GREATLEY-HIRSCH

> compute, *v*. 1.a. To determine by arithmetical or mathematical reckoning; to calculate, reckon, count. In later use chiefly: to ascertain by a relatively complex calculation or procedure, typically using a computer or calculating machine.
>
> (*OED*)

For over 400 years, our first contact with Shakespeare as children and adults has been on the page – in books printed in an infinite variety of languages, formats, shapes and sizes – or on the stage, in theatres great and small. But times have changed. Today, millions of people around the world access Shakespeare's works online – on their smartphones, tablets and computers – many for the first time. For an ever-increasing number of users, digital technologies are shaping how we experience Shakespeare and engage with his works, as both consumers and producers. Large-scale digitization projects and the proliferation of databases are making primary and secondary materials hitherto more readily available to those critics fortunate enough to have access. These digital tools are introducing new forms of criticism while also enabling traditional analysis at unprecedented scale, speed and efficiency; at the same time, digital publishing is opening up new platforms for the dissemination of scholarship.

While digital Shakespeare content is a proving rich area for critical study in its own right,[1] media historians remind us that the difficulty for scholars is that any 'new media' does not stay new for long. Unlike the codex, which has remained a relatively stable format over the course of its history, the web is constantly evolving. Attempts to survey the impact of new media on Shakespeare study thus become quickly dated, as each new wave of technological innovation renders past conclusions obsolete.[2] Likewise, the exponential growth rate of digital Shakespeare content makes the task of accurately cataloguing new material a near impossibility. Perhaps paradoxically, the rhetoric of new media reveals the contingency of its 'newness', constructed as both radically different from and reassuringly similar to the past: we access 'pages' and navigate them by 'scrolling', for instance, but the web is neither codex nor manuscript. With this in mind, this chapter's discussion of computational studies of Shakespeare will stress both its departures from and continuities with earlier forms of quantitative criticism.

SHAKESPEARE AND/AS DATA

Scholars have counted things in Shakespeare long before the advent of the computer made the process more accurate, efficient and sophisticated. As we shall see, *what* is counted and *how* it is counted have changed, in no small part due to the affordances of computing.

The earliest application of quantitative methods to the study of Shakespeare appeared during the eighteenth century in the form of the *concordance* – an index of the words used in a text or corpus keyed to citations of the passages in which they occur. With origins tracing back to medieval scriptural exegesis and adapted only relatively recently to the study of non-biblical literature,[3] the basic function of the concordance is 'to bring together ("concord") passages of a text that illustrate the uses of a word' (Howard-Hill 1979: 4). Andrew Becket compiled the earliest concordance to Shakespeare's plays in 1787, designed, as advertised by its title, to give readers ready access to 'distinguished and parallel passages' arranged alphabetically by theme. Becket's selective index of axiomatic passages was followed by increasingly more precise concordances to the plays compiled by Samuel Ayscough (1790), Francis Twiss (1805), Mary Cowden Clarke (1844–5) and W. H. Davenport Adams (1886), as well as concordances to the poems by James Orchard Halliwell-Phillipps (1867) and Helen Kate Rogers Furness (1874). Keyed to the popular Globe edition of Shakespeare's works,[4] John Bartlett's *Complete Concordance* (1894) to the plays and poems remained the standard reference of its kind for most of the twentieth century until it was superseded by systematic computer-generated concordances prepared by T. H. Howard-Hill (1969–73) and Marvin Spevack (1968–80). Both Howard-Hill and Spevack's concordances are of a scale, accuracy and detail not seen before or since in print, and, unlike their predecessors, they provide summary statistics about the texts. Over the course of nine volumes, Spevack offers a series of interlocking concordances to the plays and poems – individually, collectively and as a complete works – with plays further concorded by character. The entry for each play, for example, begins with summary statistics about the total number of speeches, lines, word-tokens (in verse, prose and mixed contexts) and word-types, while the concordances give raw and relative frequencies for each word.

Beyond their use as hermeneutic tools, concordances were also instrumental in thoroughly debunking the myth that Shakespeare's vocabulary dwarfed those of all other English playwrights (Crystal 2008). There is some truth to it: when we tally the different words used by Shakespeare and his fellow dramatists, Shakespeare's total is indeed larger. However, since Shakespeare's plays survive in greater number than those of any other single playwright of the period, the comparison is skewed: a substantially larger canon means significantly more opportunities to use different words. A simple comparison of total word-types is therefore a misleading measure of vocabulary. A more meaningful comparison is to calculate the rates at which Shakespeare and his contemporaries introduced new words with each successive play. Counted in this way, computational studies have shown Shakespeare's practice to be decidedly average amongst his contemporaries (Craig 2011; Elliott and Valenza 2011).

Though it was amongst the earliest to be quantified, vocabulary is but one feature of Shakespeare's writing to pique the curiosity of critics. The nineteenth century witnessed a veritable gold rush for the quantitative analysis of Shakespeare's works, with investigators mining every feature for discernable and distinctive patterns that might characterize aspects of Shakespeare's style. Much of this stylometric work was conducted under the auspices of the New Shakspere Society, founded in 1873 by F. J. Furnivall to 'do honour to SHAKSPERE, to make out the succession of his plays, and thereby the growth of his mind and art' (Furnivall 1874b: 6). In the words of one of its most industrious members, F. G. Fleay, criticism of Shakespeare 'must become quantitative'; the necessary 'great step' championed by the Society was to 'cease to be empirical, and become scientific': 'if you cannot weigh, measure, [or] number your results, however you may be convinced yourself, you must not hope to convince others, or claim the position of an investigator; you are merely a guesser, a propounder of hypotheses' (Fleay 1874: 2). Although the initial goal of the Society was to determine the chronology of Shakespeare's plays 'by a very close study of the metrical and phraseological peculiarities' of his works (Furnivall 1874a: vi), its focus on countable features of verse inevitably led to 'the determination of the genuineness of the works traditionally assigned to a writer': authorship attribution, which Fleay termed 'the far more important end' of their researches (1874: 6). In his first paper before the Society, Fleay tabulated the rates of double or 'feminine' endings, pause-ended or 'stopped' lines, rhyming lines, incomplete lines and Alexandrines. On the basis of these counts, he suspected that *The Taming of the Shrew* and parts of *Henry VIII*, *Pericles*, *Timon of Athens* and the Henry VI plays were not by Shakespeare (Fleay 1874). Critical consensus on Shakespeare's collaborations has since confirmed many of Fleay's suspicions.

While the successful findings of the New Shakspere Society are tempered by many false starts and outright failures, its spirit of exploration and experimentation inspired fresh quantitative investigation long after its dissolution in 1894. Since the subject of authorship attribution is given more detailed treatment in another *Arden Shakespeare Handbook* and elsewhere,[5] my discussion of its development will be accordingly brief. With some refinements, the attribution methods pioneered by the New Shakspere Society – 'counting the frequencies of certain verse features' and 'finding parallel passages' – remained 'essentially unchanged for the next 100 years' (Egan 2017: 33). A watershed moment for stylometry was the introduction of the desktop computer in the 1970s. Machine-readable texts of Shakespeare's works were soon prepared, and if computers were able to produce accurate concordances of *all* Shakespeare's words, they could also be used to count *any* feature of Shakespeare's writing, not merely certain habits of verse. Function words,[6] too frequent to be concorded by hand and routinely excluded from computation as so-called 'stop words', emerged as an especially weighty stylistic feature, not least by virtue of their ubiquity.[7]

Where computer-generated datasets like Howard-Hill and Spevack's concordances were once only distributed in print, widespread adoption of personal computing and electronic publishing has since opened up new possibilities for data production and dissemination. An equally crucial moment for stylometry thus began in the 1990s,

when large-scale commercial databases published by Chadwyck-Healey made a significant proportion of early modern literature available in machine-readable formats on CD-ROMs: *English Poetry* (1992–5), *English Verse Drama* (1995), *English Prose Drama* (1996–7) and *Early English Prose Fiction* (1997). These and other databases were later consolidated into *Literature Online* (LION), delivered as a website, and by 2000 Chadwyck-Healey (now ProQuest) embarked on a new venture – the Text Creation Partnership (TCP) – freshly to transcribe a subset of the texts in its *Early English Books Online* (EEBO) database, launched in 1998. Between them, these large-scale digitization projects have made a significant proportion of the corpus of early modern texts available in machine-readable formats, but coverage is selective, and the transcriptions are partial and frequently marred by errors.[8]

CASE STUDY

A qualitative interpretation of a work by the close reading of selected passages is not the same thing as a systematic analysis of the work in its entirety – line by line, word by word. Likewise, we accept as normal practice the limited focus necessary for a literary history to produce a coherent narrative. In both of these examples, literary critics engage in 'data reduction' – the inclusion of some features and the exclusion of others to make sense of larger phenomena. The same approach is essential to quantitative and computational criticism, by which means an investigator constructs and tests models – representations that cannot account for all aspects of the phenomena being modelled, but which nonetheless allow for valid inferences to be made (Piper 2017; Jannidis and Flanders 2019). Since these representations focus on some features and disregard others, models always involve some information loss. The inclusion and exclusion of features is not random, however, but functional: just as histories of early modern theatre necessarily privilege certain plays, playwrights, playing companies and playhouses and exclude others to produce a coherent narrative, so too is a road map a useful model for navigating terrain because of the features selected for inclusion (e.g. roads, highways, landmarks), even if most of the information about that terrain – geographical, political, social – is lost.

In this case study, I use a computational model to explore the dialogue of Shakespeare's plays. To construct my chosen corpus, I have extracted the base text from a digital copy of the Arden *Complete Works* (1998) and removed all the prefatory and editorial matter. I have then used textual encoding to annotate or 'tag' act and scene divisions, speeches and speech prefixes, and stage directions for all thirty-eight plays.[9] By searching for patterns in the text using regular expressions,[10] I added tags to define the expanded or regularized forms of contractions, abbreviations and compound words. To check my dataset for errors and inconsistencies, I then validated[11] the documents and imported them into Intelligent Archive, a software application designed for text processing, which I used to generate complete concordances to each play.[12] Since my only concern in this case study was dialogue, all other features of the text – stage directions, literary divisions, speech prefixes and so on – were excluded from my counts.

To select features for my models and process the data, I used Intelligent Archive to count the top 100 most frequent words across the corpus, calculated as a proportion of total tokens for each play.[13] Table 1 lists the plays included in the case study, with dates of first performance and genre classifications from the revised *Annals of English Drama* (Harbage and Schoenbaum 1964). The result is a table with 38 rows (one for each play) and 100 columns (one for each word). By treating each of the proportions as a coordinate, we could project each play as a data-point in 100-dimensional space. This would allow us to measure the distances between plays, thereby getting a sense of their relative affinities and differences. But there is a problem: while computers can easily model spaces in 100 dimensions, human cognition is – at time of writing, at least – limited to perceiving no more than three. To work around this problem, I used a standard statistical procedure called Principal Component(s) Analysis (PCA) to reduce the dimensionality of the data.[14] PCA attempts to explain as much of the variation in a dataset as possible using as few of the variables as possible. This is accomplished mathematically by condensing multiple variables that are correlated with one another, but largely independent of others, into a smaller number of composite factors. (In this context, 'variables' are quantifiable features capable of varying in value, such as the frequency or proportion of the word *them* in different plays. 'Correlation' describes a relationship of interdependence between two or more variables, in which a change in the value of one is associated with a change in the value of the others.) The strongest composite factor – the one that accounts for the largest proportion of the total variance in the data – is called the first principal component (PC1); the second principal component (PC2) is the composite factor accounting for the greatest proportion of the remaining variance, while also being uncorrelated with the first principal component. Further principal components can be calculated in this manner, each accounting for a smaller proportion of the remaining variance in the data than the last.

Table 1 Plays in the corpus.

Play	*Genre*	*Date of first performance*
1 Henry IV	History	1597
2 Henry IV	History	1597
1 Henry VI	History	1592
2 Henry VI	History	1591
3 Henry VI	History	1591
All's Well that Ends Well	Comedy	1603
Antony and Cleopatra	Tragedy	1606
As You Like It	Comedy	1599
The Comedy of Errors	Comedy	1594
Coriolanus	Tragedy	1608

Play	*Genre*	*Date of first performance*
Cymbeline	Tragicomedy	1610
Hamlet	Tragedy	1601
Henry V	History	1599
Henry VIII	History	1613
Julius Caesar	Tragedy	1599
King John	History	1596
King Lear	Tragedy	1605
Love's Labour's Lost	Comedy	1595
Macbeth	Tragedy	1606
Measure for Measure	Comedy	1603
The Merchant of Venice	Comedy	1596
The Merry Wives of Windsor	Comedy	1597
A Midsummer Night's Dream	Comedy	1595
Much Ado about Nothing	Comedy	1598
Othello	Tragedy	1604
Pericles	Tragicomedy	1608
Richard II	History	1595
Richard III	History	1592
Romeo and Juliet	Tragedy	1595
The Taming of the Shrew	Comedy	1591
The Tempest	Comedy	1611
Timon of Athens	Tragedy	1605
Titus Andronicus	Tragedy	1592
Troilus and Cressida	Tragedy	1602
Twelfth Night	Comedy	1601
The Two Gentlemen of Verona	Comedy	1590
The Two Noble Kinsmen	Tragicomedy	1613
The Winter's Tale	Tragicomedy	1609

To analyse the data, I imported the table of word-frequency counts into R, a software environment for statistical computing, and used the built-in PCA algorithm to reduce the dimensionality of the data to the two strongest factors.[15] Treating scores on PC1 and PC2 as *x*- and *y*-coordinates, I projected each play as a data-point in two-dimensional space (Figure 3). I added labels and symbols to identify the data-points by their abbreviated title and genre, since this metadata was withheld from the PCA algorithm and played no part in its calculations. As per the legend at the top of the chart, comedies are plotted as filled circles, histories as filled triangles,

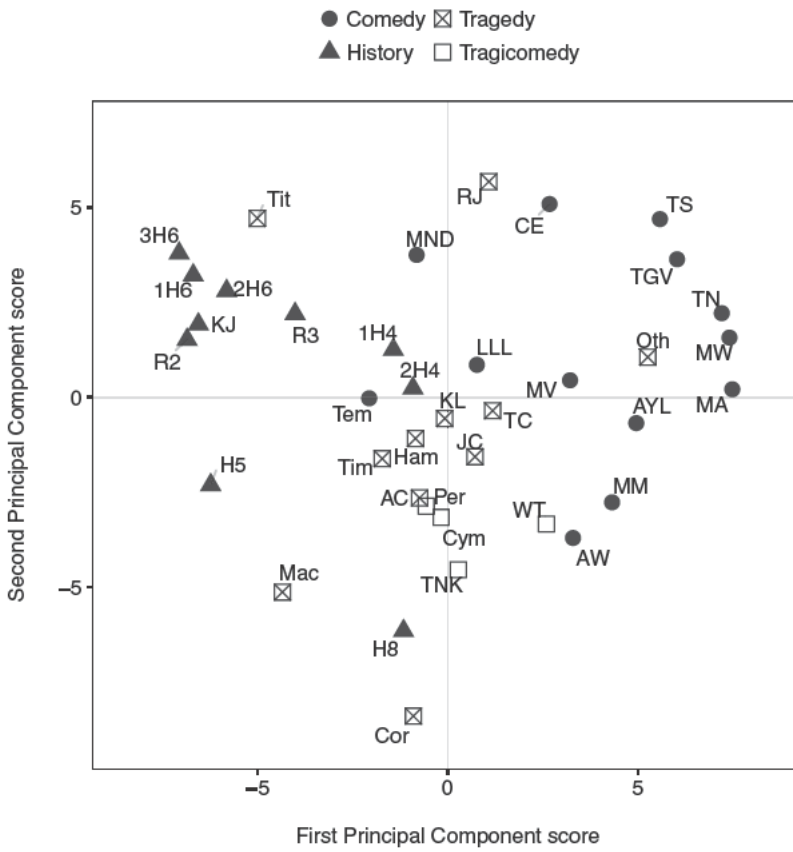

FIGURE 3 Principal Components Analysis (PCA) scatterplot.

tragedies as checked squares and tragicomedies as unfilled squares. PC1 (the *x*-axis) represents the most important latent factor in the underlying relationships between the top 100 most frequent words in the plays and PC2 (the *y*-axis) the second most important (independent) latent factor. The relative distances between data-points projected into this two-dimensional space represent degrees of affinity, so that plays with similar rates of occurrence of the top 100 most frequent words – an aspect of their linguistic profiles – cluster together, whereas dissimilar plays are plotted further apart.

When analysing plays by different playwrights of the same period, authorship consistently emerges as the strongest determining factor in the language of early modern drama (Craig 2000; 2017; Craig and Greatley-Hirsch 2017). Although several plays in my Shakespearean corpus are collaborations, the vast majority are single authored; without more pronounced, competing authorial signals, genre emerges as a stronger determining factor. We see this in Figure 3, where all of

the histories cluster together to the left of the chart, scoring negatively on PC1, whereas the comedies tend to cluster at the opposite end, mostly scoring positively on PC1; tragedies and tragicomedies are plotted towards the middle, between the histories and comedies. Conditioned by the Aristotelian model of drama, we might have expected comic and tragic drama to be the most dissimilar; however, the PCA shows that there is a greater contrast in dialogue between Shakespearean comedies and histories, which cluster at opposite ends of PC1. This confirms the findings of other studies (Craig 2004; 2008; Hope and Witmore 2004; 2010), but we need to examine the composition of each principal component before we might begin to explain these results.

Since PCA works by finding weightings for the variables to establish new composite factors or 'principal components', we can examine these weightings to find out which variables contribute the most to a given component. Using a biplot (Figure 4), we can visualize these contributions in the same two-dimensional space as the plays, with each variable represented by a vector or line projected from

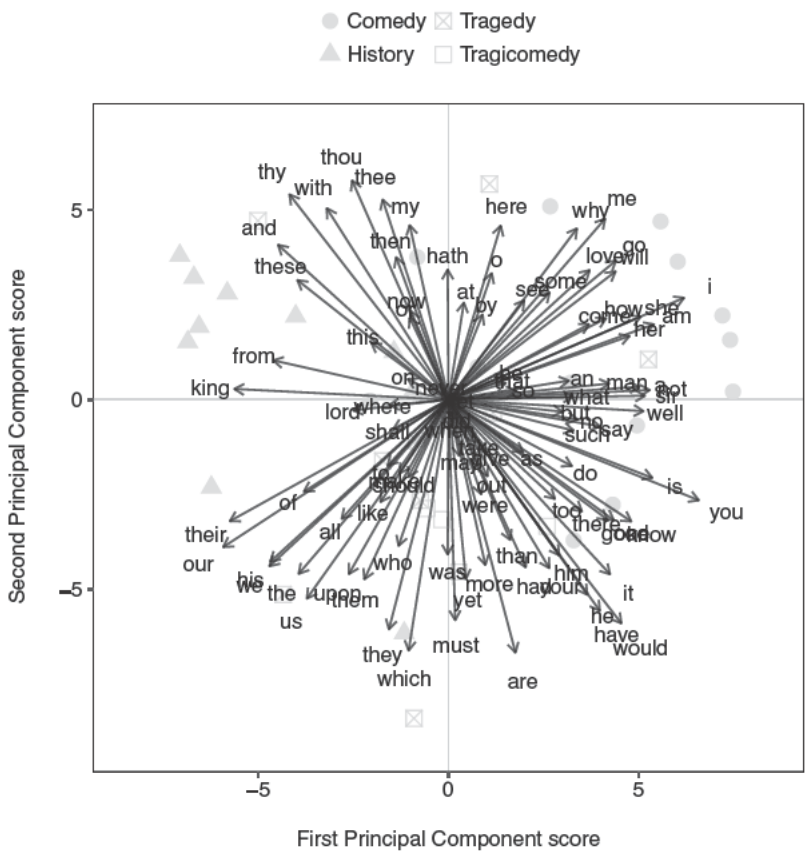

FIGURE 4 Principal Components Analysis (PCA) biplot.

the origin (where the *x*- and *y*-axes intersect). A vector's direction indicates how the variable it represents contributes to each of the principal components, and its relative length indicates the magnitude of contribution.[16] Looking along the *x*-axis first, the highest positive score is for proportions of *you*; those of *I*, *am*, *is* and *she* are also closely associated with it. These variables are strongly correlated: where one of them is notably frequent in a play, the others tend to be frequent, too, and vice versa. At the other end of the *x*-axis, *our* is the extreme, strongly correlated with words like *their*, *king*, *we* and *from*.

The pattern that emerges along PC1 appears to be a contrast between words associated with interactive dialogue and words associated with description and narration (Craig 2004). Reading the biplot (Figure 4) in relation to the earlier scatterplot (Figure 3), we find histories characterized by a higher proportion of words associated with description and narration, such as *these*, *from*, *this*, *of*, *the*, *upon*, *then* and *now*, words with strongly negative weightings on PC1. The higher proportion of certain pronouns reflects a further social dimension of the histories, in which the second person pronouns used to address individuals of subordinate social rank or a group (*thee*, *thou*, *thy*) and the titles of those in a position to use them (*king*, *lord*) are more common. Plural personal pronouns (*our*, *them*, *they*, *their*) are also weighted negatively on PC1 and more common in the histories, as are the words *we* and *us* (counts of which, in this study, conflate the royal and true plural forms). By contrast, comedies are generally marked by a higher proportion of words associated with interactive dialogue – words with strongly positive weightings on PC1, including the second person pronouns used in exchanges between social equals (*you*, *your*).[17] Whereas the lofty concerns of court and country in the histories require more frequent addresses to groups (e.g. courtiers, soldiers, subjects, nations), monarchs to themselves in the plural, narration and description, the mundane and mercantile world of the comedies is built upon direct exchanges between characters talking about themselves and one another: *I*, *me* and *you*, three words with the strongest positive weightings on PC1, predominate in the comedies. With the notable exception of *his*, all of the gendered third person pronouns are also weighted positively on PC1 (*she*, *her*, *he*, *him*), reflecting the greater centrality of women in the comedies as characters and plot devices. In this model of the plays, *pace* Linda Bamber (1982), we find comic women and not so much tragic but historic men.

Along the *y*-axis, we find two separate but related language patterns. The first is a contrast between intimate discourse (marked by the use of first person pronouns *I*, *me*, *my* and the exclamatory *O*), weighted positively on PC2, and public discourse (with more frequent recourse to plural pronouns like *their*, *our*, *we*, *us*, *them* and *they*), weighted negatively. The second is a contrast between old-fashioned and modern language: older forms (*hath*, *thy*, *thou*, *thee*) have strong positive weightings on PC2, while more modern forms (*have*, *do*, *you*, *your*) are negatively weighted. In this, Shakespeare is like his fellow dramatists: the replacement of *th*-forms with *y*-forms is 'one of the most marked developments through the period' (Craig 2008: 287). Since they are 'associated with archaic or formal language' and 'used in contexts favouring linguistic conservative', it is unsurprising that

th-forms are more common in the histories, where 'formal challenges and defiance are issued' and their 'archaic-religious and archaic-legal connotation' has more dramatic import (Hope 2003: 80–1). Moreover, there appears to be a relationship between a play's date of first performance (Table 1) and its score on PC2. We can express the strength of this relationship by calculating its correlation coefficient, a standard statistical measure given as a value between −1 (for a perfect negative correlation) and 1 (for a perfect positive correlation), where 0 indicates no relationship between the variables. The correlation coefficient between a play's date of first performance and its score on PC2 is −0.83, a very strong negative correlation.[18] In other words, early plays tend to score positively on PC2, whereas later plays tend to score negatively.

There are exceptions to this chronological pattern: for example, we might expect *The Tempest*, as one of the last plays to be written, to feature a higher proportion of modern forms and therefore score much lower on PC2. While most of the characters in *The Tempest* prefer *y*-forms to *th*-forms, in keeping with the play's date of composition, Prospero and Caliban – the two largest parts in terms of dialogue – prefer *th*-forms to *y*-forms. In their speech, Prospero and Caliban use *thee* and *thou* roughly three times as often as they use *you* and *your*. For Prospero, the *th*-forms are 'a handy shorthand for his miniature patriarchy, a tiny kingdom more or less willingly bound to its father–ruler' (Craig 2008: 287). For Caliban, the *th*-forms are an expression of resentment and defiance, as in 'This island's mine by Sycorax, my mother, / Which *thou* tak'st from me. When *thou* cam'st first / *Thou* strok'st me and made much of me' (*Tem* 1.2.332–4, emphasis added), and a reflection of his coarseness (Byrne 1936: 137–40).

Plays of the same genre typically cluster together (Figure 3), sharing similar proportions of words weighted on the first and second principal components (Figure 4). However, there are some interesting outliers and anomalies. *Othello*, for instance, scores the highest on PC1 of any tragedy and is plotted with the cluster of comedies. Hope and Witmore made the same observation in their study: built 'on structures that would ordinarily be employed in comedy' to 'heighten[] the emotional effect of down-turn' as the play reaches its tragic conclusion, *Othello*'s language profile is 'not true to type' and more closely aligns with comedies than with other tragedies (2010: 374, 376ff.). These findings, also derived computationally, confirm Susan Snyder's qualitative study of *Othello*'s 'comic matrix' (2002: 29–45). At the other extreme, *Macbeth* and *Titus Andronicus* score the lowest of the tragedies on PC1 and are plotted with the histories. As with *Othello*'s 'comic matrix', critics have noted affinities between these plays and the histories. E. M. Tillyard concluded *Shakespeare's History Plays* with a short chapter on *Macbeth*, described as not only 'the last of the great tragedies' but also 'the epilogue of the Histories' (1944: 315); for Stanley Cavell, *Macbeth* belongs 'as much with Shakespearean histories as with the tragedies' (1992: 1). *Titus Andronicus*, on the other hand, was Shakespeare's earliest effort in a genre to which he would not return until much later: 'If we set aside for the moment *Titus Andronicus* as a revenge tragedy in a genre that Shakespeare chose not to pursue further during his early years in London,' David Bevington concludes, 'we can say that Shakespeare

began his career as a dramatist chiefly as a deviser of romantic comedies and English history plays' (2011: 85).[19]

'THERE IS MEASURE IN EVERYTHING' (*MA* 2.1.65)

Given data enough and time to develop the necessary methods, quantitative criticism promises to shed light on many unresolved questions about Shakespeare's canon, chronology, sources and style, as well as his relationship to his contemporaries. Although such investigations cannot be truly exhaustive until accurate machine-readable texts for the entire corpus of early modern texts extant in print or manuscript are produced, meaningful results can still be derived using the data currently available. As the amount of data steadily grows and computational methods are developed or adapted, the range of critical applications expands accordingly.

Where scholars once had to rely upon their reading and memory, the task of locating textual parallels for the study of Shakespearean authorship, sources, lost plays and editorial cruxes is increasingly accomplished through the use of 'string matching' and 'sequence alignment' algorithms, designed to find exact and approximate matches with a given 'string' or sequence of characters or words across a corpus of texts (Steggle 2014; 2015; Greatley-Hirsch and Johnson 2018). Similar computational methods have also been used to collate and quantify textual variation in Shakespearean texts (Widmann 1971; 1973; Horton 1994). Computational analysis of Shakespeare's plays, in isolation and in relation to the works of his peers, has revealed fascinating insights into the language of genre (Craig 1991; 2017; Hope and Witmore 2004; 2010; Witmore, Hope and Gleicher 2016; Craig and Greatley-Hirsch 2017), Shakespeare's 'late' style (Hope and Witmore 2007; 2014), characterization (Craig 2008; Culpeper 2014; Algee-Hewitt 2017; Craig and Greatley-Hirsch 2017) and use of soliloquies and asides (Nordlund 2014).

Other studies have employed quantitative methods to challenge claims made about repertory company styles (Basu, Hope and Witmore 2017; Craig and Greatley-Hirsch 2017), to discern generic patterns in the distribution of stage props (Teague 1991; Bruster 2002; Craig and Greatley-Hirsch 2017), to map the Elizabethan book trade (Farmer and Lesser 2013), to identify statistical patterns in dramatic verse (Jackson 2002; Tarlinskaja 2014; Bruster and Smith 2016; Taylor and Loughnane 2017) and to track variation across foreign-language theatrical translations (Cheesman 2015; Geng et al. 2015). As for Shakespeare's poetry, critics have explored the use of machine learning techniques to locate the *volta* in the Sonnets (Katajamäki, Honkela and Kohonen 2005) and to classify and count rhetorical figures (Bradley and Ullyot 2018).

It was inevitable, perhaps, that Shakespeare criticism would itself become the subject of quantitative analysis ere long. Using computational methods to analyse data from bibliographies, catalogues and reference works, investigators have studied the editorial treatment of Shakespeare in relation to his fellow dramatists (Hirsch 2011; Lopez 2014), identified critical trends in Shakespeare scholarship and publishing (Estill, Klyve and Bridal 2015), reassessed rates of dramatic collaboration and composition (Brown 2017; Loughnane forthcoming) and explored the early modern social networks of the professional London theatres (Basu, Hope and

Witmore 2017), the book trade (Farmer and Lesser 2013; Greteman 2014–present) and notable figures from the period (Warren 2012–present). Data begets analysis, and analysis in turn becomes data to be analysed further. It is only a matter of time before investigators begin to conduct quantitative Shakespeare meta-criticism.

NOTES

I wish to thank Hugh Craig, Gabriel Egan and Sarah Neville for their invaluable feedback on this chapter, which is dedicated to the memory of two gentle giants: David Bevington and John Burrows.

1. Representative examples include essays in the following recent collections: Rowe (2010), Carson and Kirwan (2014), Hirsch and Craig (2014b), Estill, Jakacki and Ullyot (2016), Jenstad, Kaethler and Roberts-Smith (2018) and O'Neill (2019).
2. Representative examples of scholarly overviews include Lancashire (2002), Best (2011), Hirsch and Craig (2014a), O'Neill (2014), Greatley-Hirsch and Best (2017) and Wilson (2018).
3. On the history and utility of concordances, see Howard-Hill (1979) and Higdon (2003).
4. On the popularity of the Globe edition, see Murphy (2003).
5. For a general introduction to authorship attribution in theory and practice, see Love (2002); for more recent overviews of computational methods, see Juola (2006) and Luyckx (2010). Egan (2017) provides a detailed historical survey of Shakespeare attribution study; see also Hope (1994), Sharpe (2013) and Taylor and Loughnane (2017).
6. Function words are those expressing a grammatical relationship or classification or clarifying syntactic relationships.
7. Burrows (1987) pioneered computational stylistics and the analysis of function words; see also Craig and Greatley-Hirsch (2017).
8. For a history and critique of these resources, see Kichuk (2007), Gants and Hailey (2008) and Gadd (2009).
9. Each play was represented by its own document formatted in XML (eXtensible Markup Language) using tags conforming to the guidelines of the TEI (Text Encoding Initiative).
10. A regular expression is an algebraic formula describing a pattern to be searched. A familiar example of a regular expression is the wildcard notation, whereby a search for *text** returns matches for *text* as well as *textile*, *textiles*, *texts*, *textual*, *textuality* and so on. Regular expressions can be used to search for more sophisticated patterns. For example, a search of *Antony and Cleopatra* using the regular expression *[a-z]'t\r* finds *was't*, *do't*, *unto't* and *into't* at the ends of lines 2.6.14, 4.1.16, 4.14.17 and 4.14.101.
11. XML provides a means for representing the abstract structure of an ideal document against which other documents can be checked. A 'valid' XML document is one that conforms to this ideal.
12. This process allows me to spot, for example, some rogue compound contractions I had failed to regularize on the first pass because the regular expressions I initially used did not make allowance for extraneous spacing between *i'* and *th'*.

13. In order of frequency, these words are: *the*, *and*, *I*, *to*, *of*, *you*, *a*, *is*, *my*, *that*, *in*, *it*, *not*, *me*, *for*, *will*, *with*, *be*, *your*, *he*, *this*, *his*, *but*, *have*, *as*, *thou*, *him*, *so*, *what*, *her*, *do*, *thy*, *we*, *no*, *all*, *by*, *shall*, *if*, *are*, *our*, *thee*, *on*, *good*, *now*, *lord*, *from*, *sir*, *come*, *she*, *would*, *they*, *was*, *at*, *let*, *or*, *here*, *more*, *which*, *there*, *am*, *O*, *well*, *how*, *then*, *them*, *their*, *us*, *when*, *love*, *hath*, *than*, *man*, *upon*, *one*, *were*, *go*, *like*, *know*, *may*, *say*, *make*, *did*, *yet*, *should*, *must*, *an*, *why*, *see*, *had*, *such*, *out*, *give*, *where*, *king*, *these*, *who*, *some*, *never*, *too* and *take*.
14. Any college textbook on multivariate statistics will provide a more detailed discussion of PCA than my chapter allows. For a gentler introduction, see Alt 1990: 48–80.
15. PC1 accounted for 18.15% of the total variance in the data and PC2 for 11.05% of the remaining; combined, PC1 and PC2 explain 29.20% of the total variance.
16. Since they are scaled to fit the biplot, only the direction and relative lengths of the vectors matter; the precise distances between the heads of the vectors and the data-points representing the plays are meaningless.
17. On pronouns as genre markers in Shakespeare, see Brainerd (1979), Craig (1991), Hope (1994) and Busse (2002).
18. Any college textbook on statistics will explain the correlation coefficient and its formula in greater detail than space allows here. For my results, I used the CORREL function in Microsoft Excel.
19. Elsewhere, Bevington has gone so far as to characterize *Titus Andronicus* as 'a fanciful history play with a deep interest in the tragic consequences of civil conflict' (2008: 42).

REFERENCES

Algee-Hewitt, M. (2017), 'Distributed Character: Quantitative Models of the English Stage, 1550–1900', *New Literary History*, 48: 751–82.

Alt, M. (1990), *Exploring Hyperspace: A Non-Mathematical Explanation of Multivariate Analysis*, Maidenhead: McGraw-Hill.

Bamber, L. (1982), *Comic Women, Tragic Men: Gender and Genre in Shakespeare*, Stanford: Stanford University Press.

Basu, A., J. Hope and M. Witmore (2017), 'The Professional and Linguistic Communities of Early Modern Dramatists', in R. D. Sell, A. W. Johnson and H. Wilcox (eds), *Community-Making in Early Stuart Theaters: Stage and Audience*, 63–94, New York: Routledge.

Best, M. (2011), 'Shakespeare on the Internet and in Digital Media', in M. T. Burnett, A. Streete and R. Wray (eds), *The Edinburgh Companion to Shakespeare and the Arts*, 558–76, Edinburgh: Edinburgh University Press.

Bevington, D. (2008), *Shakespeare's Ideas: More Things in Heaven and Earth*, Malden: Wiley-Blackwell.

Bevington, D. (2011), 'Shakespeare's Development of Theatrical Genres: Genre as Adaptation in the Comedies and Histories', in A. R. Guneratne (ed.), *Shakespeare and Genre*, 85–99, New York: Palgrave.

Bradley, A. J. and M. Ullyot (2018), 'Machines and Humans, Schemes and Tropes', *Early Modern Literary Studies*, 20: 3.1–16.

Brainerd, B. (1979), 'Pronouns and Genre in Shakespeare's Drama', *Computers and the Humanities*, 13: 3–16.
Brown, P. (2017), 'Early Modern Theatre People and Their Social Networks', PhD diss., De Montfort University, Leicester.
Bruster, D. (2002), 'The Dramatic Life of Objects in the Early Modern English Theater', in J. G. Harris and N. Korda (eds), *Staged Properties in Early Modern English Drama*, 67–96, Cambridge: Cambridge University Press.
Bruster, D. and G. Smith (2016), 'A New Chronology for Shakespeare's Plays', *Digital Scholarship in the Humanities*, 31: 301–20.
Burrows, J. F. (1987), *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*, Oxford: Clarendon.
Busse, U. (2002), *Linguistic Variation in the Shakespeare Corpus: Morpho-Syntactic Variability of Second Person Pronouns*, Amsterdam: John Benjamins.
Byrne, Sister St G. (1936), *Shakespeare's Use of the Pronoun of Address*, Washington: Catholic University of America.
Carson, C. and P. Kirwan, eds (2014), *Shakespeare and the Digital World: Redefining Scholarship and Practice*, Cambridge: Cambridge University Press.
Cavell, S. (1992), 'Macbeth Appalled (I)', *Raritan*, 12: 1–15.
Cheesman, T. (2015), 'Reading Originals by the Light of Translations', *Shakespeare Survey*, 68: 87–98.
Craig, H. (1991), 'Plural Pronouns in Roman Plays by Shakespeare and Jonson', *Literary and Linguistic Computing*, 6: 180–6.
Craig, H. (2000), 'Is the Author Really Dead? An Empirical Study of Authorship in English Renaissance Drama', *Empirical Studies of the Arts*, 18: 119–34.
Craig, H. (2004), 'Stylistic Analysis and Authorship Studies', in S. Schreibman, R. Siemens and J. Unsworth (eds), *A Companion to Digital Humanities*, 273–88, Malden: Blackwell.
Craig, H. (2008), '"Speak, That I May See Thee": Shakespeare Characters and Common Words', *Shakespeare Survey*, 61: 281–8.
Craig, H. (2011), 'Shakespeare's Vocabulary: Myth and Reality', *Shakespeare Quarterly*, 62: 53–74.
Craig, H. (2017), 'Authorial Attribution and Shakespearean Variety: Genre, Form, and Chronology', *Shakespeare Survey*, 70: 154–64.
Craig, H. and B. Greatley-Hirsch (2017), *Style, Computers, and Early Modern Drama: Beyond Authorship*, Cambridge: Cambridge University Press.
Crystal, D. (2008), *'Think on My Words': Exploring Shakespeare's Language*, Cambridge: Cambridge University Press.
Culpeper, J. (2014), 'Keywords and Characterization: An Analysis of Six Characters in *Romeo and Juliet*', in D. L. Hoover, J. Culpeper and K. O'Halloran (eds), *Digital Literary Studies: Corpus Approaches to Poetry, Prose and Drama*, 9–34, New York: Routledge.
Egan, G. (2017), 'A History of Shakespearean Authorship Attribution', in G. Taylor and G. Egan (eds), *The New Oxford Shakespeare: Authorship Companion*, 27–47, Oxford: Oxford University Press.

Elliott, W. E. Y. and R. J. Valenza (2011), 'Shakespeare's Vocabulary: Did It Dwarf All Others?', in M. Ravassat and J. Culpeper (eds), *Stylistics and Shakespeare's Language: Transdisciplinary Approaches*, 34–57, London: Continuum.

Estill, L., D. K. Jakacki and M. Ullyot, eds (2016), *Early Modern Studies after the Digital Turn*, Toronto: Iter and ACRMS.

Estill, L., D. Klyve and K. Bridal (2015), '"Spare Your Arithmatic, Never Count the Turns": A Statistical Analysis of Writing about Shakespeare, 1960–2010', *Shakespeare Quarterly*, 66: 1–28.

Farmer, A. B. and Z. Lesser (2013), 'What Is Print Popularity? A Map of the Elizabethan Book Trade', in A. Kesson and E. Smith (eds), *The Elizabethan Top Ten: Defining Print Popularity in Early Modern England*, 19–54, Farnham: Ashgate.

Fleay, F. G. (1874), 'On Metrical Tests as Applied to Dramatic Poetry', *Transactions of the New Shakspere Society*, 1: 1–16.

Furnivall, F. J. (1874a), 'Director's Opening Speech', *Transactions of the New Shakspere Society*, 1: v–xi.

Furnivall, F. J. (1874b), 'Founder's Prospectus of the New Shakspere Society', *Transactions of the New Shakspere Society*, 1: 6–10.

Gadd, I. (2009), 'The Use and Misuse of *Early English Books Online*', *Literature Compass*, 6: 680–92.

Gants, D. and R. C. Hailey (2008), 'Renaissance Studies and New Technologies: A Collection of "Electronic Texts"', in W. R. Bowen and R. G. Siemens (eds), *New Technologies and Renaissance Studies*, 73–92, Tempe: MRTS.

Geng, Z., T. Cheesman, R. S. Laramee, K. Flanagan and S. Thiel (2015), 'ShakerVis: Visual Analysis of Segment Variation of German Translations of Shakespeare's *Othello*', *Information Visualization*, 14: 273–88.

Greatley-Hirsch, B. and M. Best (2017), '"Within This Wooden [2.]O": Shakespeare and New Media in the Digital Age', in J. L. Levenson and R. Ormsby (eds), *The Shakespearean World*, 443–62, London: Routledge.

Greatley-Hirsch, B. and L. Johnson (2018), 'Shakespeare Source Study in the Age of Google: Revisiting Greenblatt's Elephants and Horatio's Ground', in D. A. Britton and M. Walter (eds), *Rethinking Shakespeare Source Study: Audiences, Authors, and Digital Technologies*, 253–78, New York: Routledge.

Greteman, B., dir. (2014–), *Shakeosphere*. Available online: https://shakeosphere.lib.uiowa.edu/ (accessed 1 July 2019).

Harbage, A. and S. Schoenbaum (1964), *Annals of English Drama, 975–1700*, 2nd edn, Philadelphia: University of Pennsylvania Press.

Higdon, D. L. (2003), 'The Concordance: Mere Index or Needful Census?', *Text*, 15: 51–68.

Hirsch, B. D. (2011), 'The Kingdom Has Been Digitized: Electronic Editions of Renaissance Drama and the Long Shadows of Shakespeare and Print', *Literature Compass*, 8: 568–91.

Hirsch, B. D. and H. Craig (2014a), '"Mingled Yarn": The State of Computing in Shakespeare 2.0', *The Shakespearean International Yearbook*, 14: 3–35.

Hirsch, B. D. and H. Craig, eds (2014b), 'Digital Shakespeares: Innovations, Interventions, Mediations', special section of *The Shakespearean International Yearbook*, 14.

Hope, J. (1994), *The Authorship of Shakespeare's Plays: A Socio-Linguistic Study*, Cambridge: Cambridge University Press.
Hope, J. (2003), *Shakespeare's Grammar*, Arden Shakespeare, London: Bloomsbury.
Hope, J. and M. Witmore (2004), 'The Very Large Textual Object: A Prosthetic Reading of Shakespeare', *Early Modern Literary Studies*, 9: 1–36.
Hope, J. and M. Witmore (2007), 'Shakespeare by the Numbers: On the Linguistic Texture of the Late Plays', in S. Mukherji and R. Lyne (eds), *Early Modern Tragicomedy*, 133–53, Woodbridge: Brewer.
Hope, J. and M. Witmore (2010), 'The Hundredth Psalm to the Tune of "Green Sleeves": Digital Approaches to Shakespeare's Language of Genre', *Shakespeare Quarterly*, 61: 357–90.
Hope, J. and M. Witmore (2014), 'Quantification and the Language of Later Shakespeare', *Actes des congrès de la Société française Shakespeare*, 31: 123–49.
Horton, T. B. (1994), 'Sequence Comparison and Old-Spelling Texts', *Research in Humanities Computing*, 2: 89–110.
Howard-Hill, T. H. (1969–73), *Oxford Shakespeare Concordances*, 37 vols, Oxford: Clarendon.
Howard-Hill, T. H. (1979), *Literary Concordances: A Guide to the Preparation of Manual and Computer Concordances*, Oxford: Pergamon.
Jackson, M. P. (2002), 'Pause Patterns in Shakespeare's Verse: Canon and Chronology', *Literary and Linguistic Computing*, 17: 37–46.
Jannidis, F. and J. Flanders (2019), 'A Gentle Introduction to Data Modeling', in J. Flanders and F. Jannidis (eds), *The Shape of Data in Digital Humanities: Modeling Texts and Text-Based Resources*, 26–95, New York: Routledge.
Jenstad, J., M. Kaethler and J. Roberts-Smith, eds (2018), *Shakespeare's Language in Digital Media: Old Words, New Tools*, London: Routledge.
Juola, P. (2006), 'Authorship Attribution', *Foundations and Trends in Information Retrieval*, 1: 233–334.
Katajamäki, S., T. Honkela and O. Kohonen (2005), 'In Search for Volta: Statistical Analysis of Word Patterns in Shakespeare's Sonnets', *Proceedings of the International Symposium on Adaptive Models of Knowledge, Language and Cognition (AMLKC 2005)*, 44–7.
Kichuk, D. (2007), 'Metamorphosis: Remediation in *Early English Books Online* (EEBO)', *Literary and Linguistic Computing*, 22: 291–303.
Lancashire, I. (2002), 'The State of Computing in Shakespeare', *The Shakespearean International Yearbook*, 2: 89–110.
Lopez, J. (2014), *Constructing the Canon of Early Modern Drama*, Cambridge: Cambridge University Press.
Loughnane, R. (forthcoming), 'Shakespeare and the Idea of Early Authorship', in R. Loughnane and A. J. Power (eds), *Early Shakespeare, 1588–1594*, Cambridge: Cambridge University Press.
Love, H. (2002), *Attributing Authorship: An Introduction*, Cambridge: Cambridge University Press.
Luycx, K. (2010), 'Scalability Issues in Authorship Attribution', PhD diss., University of Antwerp, Antwerp.

Murphy, A. (2003), *Shakespeare in Print: A History and Chronology of Shakespeare Publishing*, Cambridge: Cambridge University Press.
Nordlund, M. (2014), 'Shakespeare's Insides: A Systematic Study of a Dramatic Device', *The Shakespearean International Yearbook*, 14: 37–56.
O'Neill, S. (2014), *Shakespeare and YouTube: New Media Forms of the Bard*, London: Bloomsbury.
O'Neill, S., ed. (2019), 'Shakespeare and Digital Humanities: New Perspectives and Future Directions', special issue of *Humanities*, 8.
Piper, A. (2017), 'Think Small: On Literary Modeling', *PMLA*, 132: 651–8.
Rowe, K., ed. (2010), 'Shakespeare and New Media', Special Issue of *Shakespeare Quarterly*, 61.
Shakespeare, W. ([1998] 2001), *The Arden Shakespeare Complete Works*, ed. R. Proudfoot, A. Thompson, and D. S. Kastan, rev. ed., London: The Arden Shakespeare.
Sharpe, W. (2013), 'Authorship and Attribution', in J. Bate and E. Rasmussen (eds), *William Shakespeare and Others: Collaborative Plays*, 641–745, New York: Palgrave.
Snyder, S. (2002), *Shakespeare: A Wayward Journey*, Cranbury: Associated University Presses.
Spevack, M. (1968–80), *A Complete and Systematic Concordance to the Works of Shakespeare*, 9 vols, Hildesheim: Georg Olms.
Steggle, M. (2014), 'The Cruces of *Measure for Measure* and EEBO-TCP', *Review of English Studies*, 65: 438–55.
Steggle, M. (2015), *Digital Humanities and the Lost Drama of Early Modern England: Ten Case Studies*, Farnham: Ashgate.
Tarlinskaja, M. (2014), *Shakespeare and the Versification of English Drama, 1561–1642*, Burlington: Ashgate.
Taylor, G. and R. Loughnane (2017), 'The Canon and Chronology of Shakespeare's Works', in G. Taylor and G. Egan (eds), *The New Oxford Shakespeare: Authorship Companion*, 417–602, Oxford: Oxford University Press.
Teague, F. (1991), *Shakespeare's Speaking Properties*, Cranbury: Associated University Presses.
Tillyard, E. M. (1944), *Shakespeare's History Plays*, London: Chatto & Windus.
Warren, C., dir. (2012–), *Six Degrees of Francis Bacon*. Available online: www.sixdegreesoffrancisbacon.com (accessed 1 July 2019).
Widmann, R. L. (1971), 'The Computer in Historical Collation: Use of the IBM 360/75 in Collating Multiple Editions of *A Midsummer Night's Dream*', in R. A. Wisbey (ed.), *The Computer in Literary and Linguistic Research*, 57–63, Cambridge: Cambridge University Press.
Widmann, R. L. (1973), 'Compositors and Editors of Shakespeare Editions', *Papers of the Bibliographical Society of America*, 67: 389–400.
Wilson, J. R. (2018), 'Shakestats: Writing about Shakespeare between the Humanities and the Social Sciences', *Early Modern Literary Studies*, 20: 4.1–38.
Witmore, M., J. Hope and M. Gleicher (2016), 'Digital Approaches to the Language of Shakespearean Tragedy', in M. Neill (ed.), *The Oxford Handbook of Shakespearean Tragedy*, 316–35, Oxford: Oxford University Press.