Introduction

> There is no such thing as a clearly defined historical field; facts are linked to other facts in all directions, and investigation merely leads to further and further questions.[1]
>
> (W. W. Greg)

> We are all disenchanted with those picaresque adventures in pseudo-causality which go under the name of literary history, those handbooks with footnotes which claim to sing of the whole but load every rift with glue.[2]
>
> (Geoffrey Hartman)

> It is generally admitted that a positivistic history of literature, treating it as if it were a collection of empirical data, can only be a history of what literature is not. At best, it would be a preliminary classification opening the way for actual literary study and, at worst, an obstacle in the way of literary understanding.[3]
>
> (Paul de Man)

It is nearly fifty years since the most recent of these remarks were published, but the fragmentary nature of the evidence – to adapt the Porter's paradox in *Macbeth* – still provokes and unprovokes attempts towards a complete, unified literary history. This problem is particularly acute in relation to the literature and drama of the early modern period, for which, as Alexander Leggatt observes, 'the full picture can never be recovered', and the aim must be to draw plausible inferences from the available evidence, so that the literary and theatrical 'life of the time can be sketched in rough outline'.[4] A second inherent difficulty with literary history is where to draw boundaries:

[1] W. W. Greg, 'Preface', in Philip Henslowe, *Henslowe's Diary*, 2 vols. (London: A. H. Bullen, 1908), II: vii.
[2] Geoffrey Hartman, 'Toward Literary History', *Daedalus* 99.2 (1970), 355.
[3] Paul de Man, 'Literary History and Literary Modernity', *Daedalus* 99.2 (1970), 401.
[4] Alexander Leggatt, 'The Companies and Actors', in Clifford Leech et al. (eds.), *The Revels History of Drama in English*, 8 vols. (London: Methuen, 1975–83), III: 99.

changes and continuities in literary history pertain to the works themselves, but also relate to all the institutions that make publication and performance possible. As such, literary history is bound up with the histories of other art forms, as well as with political, social, and technological changes.

Although the paucity of surviving historical and documentary evidence – as well as the resulting critical uncertainties and lacunae – remains an obstacle, the introduction of computing technologies and computer-aided quantitative methods in recent years has opened up possibilities to overcome some of the difficulties through new forms of evidence, critical frameworks, and vocabularies. The new methods offer comprehensiveness and evenness of attention, if at the cost of a radical narrowing of the data to those features that can be counted. A quantitative turn brings new models of literary history, combining large datasets and computerised statistical analysis to elicit otherwise hidden or only partially glimpsed patterns, which can then inform humanistic judgement and interpretation.

Some of the findings from this quantitative turn have been in areas where we might have despaired of ever getting an answer, such as the question of whether the style of the original writer survives in translations.[5] Others bring precision where we already knew the basic facts, as with the discovery that Henry James's style changed progressively with each new novel, with hardly a backward step, in the direction of his familiar late style.[6] A third variety reveals an unsuspected minuteness of patterning in literary texts, as in the distinctiveness of the speaking styles of the characters within a novel.[7]

The computer possesses some attributes of the ideal reader of narrative theory – such as the capacity to 'retain in memory and retrieve at will' the information provided by the text – while completely lacking others, such as 'the competence required to properly understand and interpret a literary work'.[8] Meanwhile empirical studies of actual readers show that they are quite unlike the ideal reader: for instance, they vary considerably over the course of reading a text in their engagement with the narrative.[9]

[5] Jan Rybicki, 'The Great Mystery of the (Almost) Invisible Translator: Stylometry in Translation', in Michael P. Oakes and Meng Ji (eds.), *Quantitative Methods in Corpus-Based Translation Studies* (Amsterdam: John Benjamins, 2012), 231–48.

[6] David L. Hoover, 'Corpus Stylistics, Stylometry, and the Styles of Henry James', *Style* 41.2 (2007), 174–203.

[7] John Burrows, *Computation into Criticism: A Study of Jane Austen and an Experiment* in Method (Oxford: Clarendon Press, 1987).

[8] Peter Dixon and Marisa Bortolussi, 'Fluctuations in Literary Reading: The Neglected Dimension of Time', in Lisa Zunshine (ed.), *The Oxford Handbook of Cognitive Literary Studies* (Oxford University Press, 2015), 543, 541.

[9] Dixon and Bortolussi, 'Fluctuations', 542–53.

The computer can read more, and more evenly, than any human reader, and this paradoxical situation – perfect evenness, unlimited memory, entire lack of comprehension – brings a capacity to offer results which might not be anticipated, which diverge from the conclusions of both ideal and actual readers, but which can be directly and completely related to the details of the text.

As the long history of the concordance demonstrates,[10] scholars since the Middle Ages have been both able and interested to count features of language in texts. To do so on a large scale, however, was simply impractical until the advent of computing, which, alongside the increasing availability of machine-readable transcriptions, as well as software applications for their processing and analysis, now allows us to consider more complex forms of evidence, such as multivariate patterns of language use and word distribution. Moreover, the application of quantitative methods and statistical reasoning to literary studies challenges scholars to situate their findings within degrees of probability, rather than making simple declarations of fact. Unexpectedly, perhaps, the quantitative approach leads to measured uncertainty rather than absolute findings. The methods foreground the possibility that a pattern is the result of chance, for instance. Tests for statistical significance frame the result: is it the sort of difference that we could expect to appear now and then, even when there is no genuine underlying contrast, or, on the other hand, is it so marked and persistent that it would take hundreds of trials of random data to come up with something similar – or thousands, or millions? (Admittedly, the authority we tend to give to numbers may obscure the other confounding possibility, that there is a hidden factor behind the results other than the one which has been targeted or detected, and investigators need to remind themselves constantly of this aspect.)

Our book does not seek to establish new certainties, but to present principled generalisations about literary history, drawn from datasets that are inevitably incomplete but nonetheless designed to bear specifically on the question at hand, and which would certainly exceed the possible span of an unaided reader. It is based on a conviction that Paul de Man, quoted above, was too pessimistic about quantitative studies of literature, which, in the decades since his remarks were published, have made valuable contributions to the discipline, by resolving some specific categorical questions like authorship and dating, and by providing some well-founded wider stylistic contexts in which particular examples can be placed. Quantitative

[10] David Leon Higdon, 'The Concordance: Mere Index or Needful Census?', *Text* 15 (2003), 51–68.

methods have distinct advantages in comprehensiveness, in giving equal attention to every instance of a feature under investigation, and in the use of well-established statistical techniques to counter potential biases. We do not suggest for a moment, however, that quantitative study can replace qualitative study. Any numerical analysis depends on previous scholarship to define the problem and the associated set of texts and variables for study, and its findings must then be interpreted in the light of a wider disciplinary understanding.

Style and Stylistics

Style is the key enabling concept in these studies. It implies a common factor linking a number of local instances and a number of markers. Writing in 1589, George Puttenham captures this sense of continuity and extent in *The Art of English Poesy*:

> Style is a constant and continual phrase or tenor of speaking and writing, extending to the whole tale or process of the poem or history, and not properly to any piece or member of a tale, but is of words, speeches, and sentences together a certain contrived form and quality, many times natural to the writer, many times his peculiar by election and art, and such as either he keepeth by skill or holdeth on by ignorance, and will not or peradventure cannot easily alter into any other.[11]

From the perspective of linguistics, style covers patterned variation within a language, an alternative to the usual focus of the discipline on the underlying syntactical rules and lexical resources of a language as a whole. For stylistics, style is a marked concentration or foregrounded under-representation of some linguistic items, and thus always a matter of relativities:

> Style is concerned with frequencies of linguistic items in a given context, and thus with *contextual* probabilities. To measure the style of a passage, the frequencies of its linguistic items of different levels must be compared with the corresponding features in another text or corpus which is regarded as a norm and which has a definite relationship with this passage. For the stylistic analysis of one of Pope's poems, for instance, norms with varying contextual relationships include English eighteenth-century poetry, the corpus of Pope's work, all poems written in English in rhymed pentameter couplets, or, for greater contrast as well as comparison, the poetry of Wordsworth.

[11] George Puttenham, *The Art of English Poesy*, ed. Frank Whigham and Wayne A. Rebhorn (Ithaca: Cornell University Press, 2007), 3.5 (233).

> Contextually distant norms would be, e.g. Gray's *Anatomy* or the London Telephone Directory of 1960.[12]

In this strictly operational sense, it is possible for a given group of language samples to have a neutral style – that is, not to have any marked differences from a comparison set. Style is also something that is perceived by readers or hearers, who compare what they are reading or hearing with acquired standards for the genre.

When styles are detected in the quantitative way Nils Erik Enkvist describes above, their make-up can be precisely specified in terms of markers and frequencies. The researcher usually wants to go beyond this to a more qualitative description, but this is much more impressionistic. In this there is the risk of committing what Ernst H. Gombrich calls the 'physiognomic fallacy', that is, the mistaken belief that styles must have the unity of a human face, a singular combination of features adding up to a naturally occurring, recognisable, separately existing entity.[13] We can present a series of feature frequencies that have high or low correlations,[14] and claim that they represent the style of a given set of samples, but must acknowledge that these are only an arbitrary selection of the possible features and thus only one of the many possible views of the set and its contrast with other comparable sets. A further step, derided by Stanley Fish and exemplified in a recent article on *Double Falsehood*, is to go beyond metaphorical physiognomy to infer actual, personal, psychological characteristics from language features in a work.[15]

The interpreter has to avoid reckless leaps into fanciful interpretation but also do justice to a sense that, though quantitatively derived, and thus arbitrary in origin, some combinations of features can be associated with the intuitive experience of readers and can crystallise otherwise half-formed

[12] Nils Erik Enkvist, 'On Defining Style: An Essay on Applied Linguistics', in Nils Erik Enkvist, John Spencer, and Michael Gregory (eds.), *Linguistics and Style* (Oxford University Press, 1964), 29.

[13] Ernst H. Gombrich, 'Style', in David L. Sills (ed.), *International Encyclopedia of the Social Sciences*, 18 vols. (New York: Macmillan, 1968–79), xv: 359.

[14] In statistics, 'correlation' describes and measures the strength (low to high) and direction (positive or negative) of the association between two sets of counts. In meteorology these counts might be rainfall and temperature; in computational stylistics, they might be frequencies of the words *you* and *thou*. A positive correlation indicates the extent to which those sets of counts or 'variables' increase or decrease in parallel, whereas a negative correlation indicates the extent to which one variable increases as the other decreases.

[15] Stanley E. Fish, 'What Is Stylistics and Why Are They Saying Such Terrible Things about It?', in Seymour Chatman (ed.), *Approaches to Poetics: Selected Papers from the English Institute* (New York: Columbia University Press, 1973), 109–52; Ryan L. Boyd and James W. Pennebaker, 'Did Shakespeare Write *Double Falsehood*? Identifying Individuals by Creating Psychological Signatures with Text Analysis', *Psychological Science* 26.5 (2015), 570–82.

impressions about distinctive strands in language use. The ancient rhetoricians provide models for identifying styles in language in this way. They identify and analyse archaic versus modern and high versus low styles in particular.[16]

It is worth noting that we need to use the term 'style' without recourse to the idea of choice, which may underpin a more specific art-historical use.[17] The notion of 'style' we employ may not have been apparent either to practitioners or to consumers.

Classification and Description

Quantitative analysis of literary language can be applied to *classification* or *description*. Classification involves the principled categorisation of texts into discrete classes on the basis of established criteria (e.g. author, genre, period, and so on), whereas description involves analysis of numerical patterns to generalise (e.g. about the change in a given genre over time) or to reveal latent aspects of texts (e.g. an unexpected contrast between the speaking styles of protagonists and antagonists in novels). We adopt both approaches in this book but do not present any work in authorship attribution, which has been the commonest application of computational stylistics to date. Broadly our aim is to build on the striking advances in rigour and diversity in authorship attribution and apply similar methods to other aspects of literary history: questions about other kinds of classification – such as by repertory company, by era, by form (verse or prose) – and more descriptive generalisation, such as dialogue types and characters. Mostly our data is linguistic, profiles of word use in particular, but for one case study we look at another kind of data: the distribution of stage properties across the plays. This quantitative work can serve our discipline by arbitrating among possible answers to the more categorical questions – of which there are more than one might imagine – and by offering fresh perspectives through the power of statistics to generalise.

The application of quantitative methods to literary study is not without its critics, and the claims of 'distant reading' in particular have provoked a considerable reaction within and beyond the academy.[18] The accommodation between a scientific paradigm and humanistic approaches is fraught

[16] Gombrich, 'Style', 354. [17] Gombrich, 'Style', 353.

[18] Representative examples include Maurizio Ascari, 'The Dangers of Distant Reading: Reassessing Moretti's Approach to Literary Genres', *Genre* 47.1 (2014), 1–19; Stanley E. Fish, 'Mind Your P's and B's: The Digital Humanities and Interpretation', *Opinionator* 23 Jan. 2012; essays in Jonathan Goodwin and John Holbo (eds.), *Reading Graphs, Maps, Trees: Responses to Franco Moretti* (Anderson: Parlor Press, 2011); Stephen Marche, 'Literature Is Not Data: Against Digital Humanities', *Los Angeles*

with difficulty, as the sweeping claims from one side are met with determined resistance from the other. The central proposition of literary Darwinism – that literature cannot be understood without taking into account the insights of evolutionary biology – has met with similar opposition.[19] Our position is that computational stylistics must take its place within the disciplinary framework of literary studies and is subject to an established understanding of the limitations of certainty in interpretation, but can nevertheless provide new insights through its power of principled generalisation. This is becoming clear in the cognate area of authorship attribution. Gary Taylor, for example, has argued that quantitative attribution work on William Shakespeare's early dramatic output has provided new opportunities for interpretation. Uncertainty about whether plays are collaborative, and which sections are written by which writer, inhibits criticism. If we know this part of *Titus Andronicus* is by George Peele, and this part by Shakespeare, we can see how they relate to the careers of these playwrights and are freed from the necessity to explain inconsistencies through textual corruption or the presence of prentice work.[20]

Quantitative Studies and the Counter-Intuitive

The ideal computational-stylistic finding is counter-intuitive but highly persuasive. In some cases, familiar critical assumptions are based on analysis of an inadequate sample of the whole or overly influenced by preoccupations and bias. A combination of empirical objectivity and expanded scale of sample opens the way to reassessing, and perhaps even invalidating, these views. The new methods may also detect patterns in certain features, which, though undeniably present, were invisible to the naked eye and thus hidden from qualitative literary-critical methods.

Inertia, ideology, and fashion have an influence on what works are studied, and what conclusions are reached – in terms both of what questions are deemed worthy of attention and what answers to the questions are preferred. Quantitative methods cannot escape the dilemmas of selectivity and bias in interpretation, but they do force the researcher to articulate a problem in objective terms and to expose a given claim to a test that can go

Review of Books 28 Oct. 2012; Adam Kirsch, 'Technology Is Taking over English Departments: The False Promise of the Digital Humanities', *New Republic* 2 May 2014; and Katie Trumpener, 'Paratext and Genre System: A Response to Franco Moretti', *Critical Inquiry* 36.1 (2009), 159–71.

[19] See Brian Boyd, 'Getting It All Wrong: The Proponents of Theory and Cultural Critique Could Learn a Thing or Two from Bioculture', *The American Scholar* 1 Sep. 2006; for a response, see Jonathan Kramnick, 'Against Literary Darwinism', *Critical Inquiry* 37.2 (2011), 315–47.

[20] Gary Taylor, 'The Fly Scene in *Titus*', paper presented at Shakespeare 450, Paris, 25 Apr. 2014.

either way. Experiment in literary interpretation, as elsewhere, can test and modify established assumptions.

Quantitative work provides an opportunity to be surprised: to back something other than the sentimental favourite and to reverse consensus views. It might have resonated better with early twenty-first-century disciplinary audiences if authorship had proved to be a muted aspect of style when analysed quantitatively. However, researchers in areas from ancient philosophy to contemporary mass-market romance have observed author-effects that cut across all the other groupings (e.g. genre and period) that can also be tested.[21] The authors of this book expected early modern repertory companies to have distinctive, identifiable styles, and that verse would constrain style whoever used it and whenever it was used. In fact, as detailed in the chapters that follow, these expectations – founded on a consensus of earlier scholarship – were consistently overturned by tests in which it was possible for the anticipated patterns to emerge. These examples suggest that the methods provide a way of avoiding confirmation bias, in which evidence supporting an established view tends to be favoured, and evidence tends to be interpreted so as to support a predetermined position. Often an established view has its own internal logic, seems plausible, and has the seductive appeal of opening up perspectives of special interest to the discipline at a given moment. A quantitative approach can offer a fresh start on the problem and challenge the researcher with unexpected findings, which in turn require further testing and explanation.

Quantitative language study works by concentrating on a select few of the manifested features of the text, ignoring the rest, as well as the context of the instances. This wilful blindness to all but a fraction of the signals to which readers and spectators respond allows all texts or text-segments to be put on exactly the same footing. The scale can be enlarged at will, limited only by the availability of suitably prepared text. A quantitative approach requires comparison to yield any results. A single reading only has significance in relation to another reading or to a standard derived from an *a priori* expectation. Comparison is built into the method. Practitioners are prevented from the sort of absolute, unanchored observations that treat a single instance in isolation, without reference to its context in comparable works.

[21] For example, Harold Tarrant and Terry Roberts, 'Appendix 2: Report of the Working Vocabulary of the Doubtful Dialogues', in Marguerite Johnson and Harold Tarrant (eds.), *Alcibiades and the Socratic Lover-Educator* (London: Bristol Classical Press, 2012), 223–36; and Jack Elliott, 'Patterns and Trends in Harlequin Category Romance', in Paul Longley Arthur and Katherine Bode (eds.), *Advancing Digital Humanities: Research, Methods, Theories* (New York: Palgrave, 2014), 54–67.

In a situation where we have a considerable collection of samples from a field such as early modern English drama, we can test how far the patterns of regularity go in particular cases. Some patterns we anticipate to be strong may prove weak in practice. The accumulation of readers' intuitions into a consensus on a given set of constraints – that a dramatic change in political régime was accompanied by a sea change in literary style, for example – may prove to be exaggerated: there was in fact only a minor change, or nothing detectable. Writing before and after this watershed is more variable than we thought – it is free to range and to innovate – and this fluctuation has for this purpose the same consequence as stability. That is, there is no marked and consistent contrast between texts either side of the divide. On the other hand, if there was a large change, it might be in a quite unexpected direction.

Authorship and Beyond

> People often say that it doesn't matter who wrote the works, we still have the works themselves . . . But it does matter. Utterly. To claim otherwise is to deny history, the nature of historical evidence, and also to sever from the works any understanding of the humanity and personality behind them. As people we want to know as much as possible about the artist responsible for the work. Even though we don't have as much personal information about Shakespeare of the kind we should like to have – diaries, letters, account-books – our desire to know as much as possible remains unabated. That is where the art of Shakespearian biography commences.[22]
>
> (Paul Edmondson and Stanley Wells)

Authorship has been the main focus of computational stylistics in studies of early modern drama to date. Since readers, playgoers, scholars, actors, and directors want to know who wrote the plays or parts of plays, it is unsurprising that new tools to classify texts have been applied first to questions of authorship. The findings have sometimes been controversial, but it is hard to imagine any new attribution now being made – or being persuasive – based entirely on a reader's sense of authorial affinities without any support from quantitative study. In 1968 Ernst Gombrich argued, 'For the time being, at any rate, the intuitive grasp of underlying *Gestalten* that makes the connoisseur is still far ahead of morphological analysis of style in terms of enumerable features' in the attribution of 'a painting, a piece of music, or a page of prose', but the balance would now seem to be

[22] Paul Edmondson and Stanley Wells, *Shakespeare Bites Back: Not So Anonymous* (Stratford-upon-Avon: Shakespeare Birthday Trust, 2011), 37.

reversed.[23] For the definition of authorial canons, the methods have become mainstream. *The New Oxford Shakespeare* (2016) is the first edition of the complete works to be predicated on a series of new inclusions and exclusions determined by quantitative study.

One side effect of the intense effort which has gone into specific, hotly contested questions of attribution is the discovery that authorial style is detectable in texts to a degree which surprises even traditional author-centred scholars. Such findings contradict many of the pronouncements of critics who prefer to see dispersed agency – through collaborative writing, or the influence of theatre or printing house participants in the process of transmission – as trumping the importance of the author.[24]

In our chapters, we take advantage of authorship work on particular questions and of the methods that have been developed and tested there. We also rely on the broader discovery about the pervasive authorial factor in linguistic style. We look beyond this powerful author-effect to other patterns in the plays, but the very strength of this effect means that we must always take it into account. We might notice a pattern of differentiation between a group of comedies and a group of tragedies, for instance, but if the majority of the comedies are by Jonson, and a majority of the tragedies by Shakespeare, then it is likely – based on the knowledge we have now accumulated – that the differences have more to do with authors than with genres. If we want to understand the nature and degree of important non-authorial considerations such as genre and era, then we must ensure that we account for any authorial effects. One simple way to do this is to observe the part played by a given grouping within the work of one author, or to make sure any one author does not dominate the sub-corpora we use for the tests.

Authorship is the prime example of a categorical question that is more important than we usually acknowledge. So much of our critical machinery will only function with a secure attribution, as the comparative neglect of anonymous and putatively collaborative works shows.[25] This is most obvious with a canonical writer: a bibliometric study would show that *A Funeral Elegy* was the object of extensive critical attention when it was (briefly) accepted as Shakespeare's, but now that it has been shown to belong to the

[23] Gombrich, 'Style', 360.

[24] Hugh Craig, 'Style, Statistics, and New Models of Authorship', *Early Modern Literary Studies* 15.1 (2009–10), 1–42; and Gabriel Egan, 'What Is Not Collaborative about Early Modern Drama in Performance and Print?', *Shakespeare Survey* 67 (2014), 18–28.

[25] For a recent survey of the editorial treatment of anonymous and collaborative plays, see Brett D. Hirsch, 'Moving Targets: Constructing Canons, 2013–2014', *Early Theatre* 18.1 (2015), 115–32.

canon of John Ford it has returned to the pack as just another early modern elegy of an obscure country squire.[26] As Shakespeare collaborations become clearer following quantitative study, and the attribution of sections to Shakespeare, Christopher Marlowe, Thomas Middleton, George Peele, George Wilkins, and others, they enable sharper and better-founded analysis and resolve long-standing puzzles. Questions of authorship are matters of classification and, in the absence of clinching documentary evidence, best resolved through the objective numerical analysis of style.

Chronology is another crucial form of classification enabling literary study not only to clarify individual literary careers and trajectories, but also to estimate the direction of influence, to chart movements and innovations, and to see works in synchronic contexts. Anywhere there is a firm classification, such as genre, mode, gender of author, gender of character, or theatrical company of first production, quantitative analysis has a role in determining the soundness or otherwise of the classification and, thus in turn, a role in enabling interpretation.

Where our questions relate to readily defined classes of literary works (or parts of them), the usefulness of computational stylistics is easy to see. We might wonder if two classes, readily defined by objective criteria, are in fact different in style from each other, such as between the generic categories of 'tragedy' and 'comedy'. If the ascription of a given sample to one class or another is disputed, we can seek in an objective way to distinguish between the two classes and apply this to the disputed work. In their quantitative analysis of Shakespeare's language classified by genre, for example, Jonathan Hope and Michael Witmore provide linguistic confirmation of Susan Snyder's earlier argument – based on qualitative readings of the plays – that a comic structure or 'matrix' underlies Shakespeare's tragedies, observing that *Othello* shares more stylistic affinities with Shakespeare's comedies than his other tragedies.[27]

The usefulness of quantitative methods in description is less obvious than for classification, but we believe it is considerable. This is the province of stylistics – the analysis of style on the basis of objectively observed

[26] On the controversy surrounding the Shakespearean attribution, see the individual essays by Richard Abrams, Stephen Booth, Katherine Duncan-Jones, Donald Foster, Ian Lancashire, and Stanley Wells in *Shakespeare Studies* 25 (1997), as well as Foster's essay and the rejoinders to it in *PMLA* 111.5 (1996) and (with Charles W. Hieatt et al.) 112.3 (1997). For a persuasive attribution to Ford, see G. D. Monsarrat, '*A Funeral Elegy*: Ford, W. S., and Shakespeare', *Review of English Studies* 53.210 (2002), 186–203.

[27] Jonathan Hope and Michael Witmore, 'The Hundredth Psalm to the Tune of "Green Sleeves": Digital Approaches to Shakespeare's Language of Genre', *Shakespeare Quarterly* 61.3 (2010), 357–90; Susan Snyder, *The Comic Matrix of Shakespeare's Tragedies* (Princeton University Press, 1979).

features. Though we do not attempt it here, one aspect of this would be the analysis of authorial characteristics. If authors' styles are indeed distinctive and consistent, so that it is possible to detect the author of a sample of unknown origin with some confidence, then it follows that we should be able to highlight and discuss some differentiating features. Not all the features that serve to distinguish authors will necessarily prove to be stylistically interesting – just as a fingerprint may identify an individual with a high degree of accuracy but tell us nothing about that person's behaviour or predispositions – but it is likely that some of the features will have a literary interest.[28]

John Burrows, Very Common Words, and Principal Components Analysis

Although we use a range of data sources and procedures in this book, we also keep returning to the alliance of counts of very common words on the data side and Principal Components Analysis on the processing side. To put this combination in context – to help in understanding a disciplinary world where it did not yet exist, and thus to see it in perspective as an innovation – it may be helpful to rehearse the story of how the method evolved.

In the early 1970s John Burrows, author of a 1968 book presenting a close reading of the characters and local interactions of *Emma*,[29] developed an interest in the patterns of the use of words that occurred regularly and were relatively inconspicuous but carried ideological freight in the novels of Jane Austen – such words as *elegant* and *nonsense*. However, in 1973 Stuart M. Tave published a book on exactly this topic, *Some Words of Jane Austen*.[30] Burrows decided to look elsewhere, and to examine Austen's use of still more regularly occurring words, such as pronouns and articles. Burrows noted that such words as *we* and *the* varied between Austen's novels in ways that could be related to questions of critical interest, such as their characters' speaking styles.

Such words had been counted before, as in Frederick Mosteller and David L. Wallace's influential authorship attribution study of *The*

[28] See Hugh Craig, 'Authorial Attribution and Computational Stylistics: If You Can Tell Authors Apart, Have You Learned Anything about Them?', *Literary and Linguistic Computing* 14.1 (1999), 103–13.

[29] John Burrows, *Jane Austen's 'Emma'* (Sydney University Press, 1968).

[30] Stuart M. Tave, *Some Words of Jane Austen* (University of Chicago Press, 1973).

Federalist Papers (1964),[31] but the laborious nature of counting instances of a word like *the* by hand – there are six instances in the preceding paragraph of this Introduction alone, for instance – restricted its statistical use to a few well-funded studies. In any case, it was assumed that there was little to be learned from these counts of very common words, which mostly have a purely grammatical function and in these cases are known as 'function words'. In Italian they are known as *parole vuote*, that is, 'empty words'. In computer science they appear on most lists of 'stop words', that is, words to be ignored by the software.

Burrows was the first to see the potential of these words for literary analysis. By the 1980s he was well advanced with studies based on writers' use of very common words, and his 1987 book, *Computation into Criticism*, is a programmatic challenge to the orthodoxy about their role in literary meaning. He notes that linguists, concordance-makers, and lexicographers continue to hold that these words are used at stable rates and carry little, if any, stylistic significance. He comments on the generally tacit understanding:

> that, within the verbal universe of any novel, the very common words constitute a largely inert medium while all the real activity emanates from more visible and more energetic bodies. The falsity of any such assumption, the inappropriateness of any such model of a verbal universe, will be established in the course of the following discussion; and the far-reaching consequences that flow from the attempt to find a better model will be seen on every side. The neglected third, two-fifths, or half of our material has light of its own to shed on the meaning of one novel or another; on subtle relationships between narrative and dialogue, character and character; on less direct and less limited comparisons between novels and between novelists; and ultimately on the very processes of reading itself.[32]

With the computer, counting these words is a trivial matter, even in the longest text. They are easy for the machine to recognise, given that in modernised text at least they are separated from their neighbours by spaces or punctuation. Once an electronic text has been created, compiling tables of word frequencies is a simple exercise.

> Computer-based concordances, supported by statistical analysis, now make it possible to enter hitherto inaccessible regions of the language, regions where, to take an extreme case, more than 26,000 instances of 'the' in Jane

[31] Frederick Mosteller and David L. Wallace, *Inference and Disputed Authorship: The Federalist Papers* (Reading: Addison-Wesley, 1964).
[32] Burrows, *Computation*, 2.

> Austen's novels defy the most accurate memory and the finest powers of discrimination and where there is diversity enough within a single novel to cast doubt on arguments based on supposedly typical specimens of Jane Austen's prose.[33]

In Cambridge, on sabbatical leave from his native Australia in 1979–80, Burrows had discussed his interest in these abundant but unevenly distributed lexical items with the Director of the University of Cambridge Computer Laboratory, Nicholas (Nick) M. Maclaren. Maclaren suggested using 'eigen-analysis', a technique for finding underlying patterns in a table of counts of multiple variables in multiple observations. This procedure was invented by Karl Pearson at the beginning of the twentieth century and again (independently) by Harold Hotelling in the 1930s.[34] Hotelling called it 'Principal Components Analysis' or 'PCA', and this name has been generally adopted.

PCA is designed to be a way of making a table of multiple observations for multiple variables comprehensible. If we take the sixty retail and hospitality businesses operating in the central business district of a small city, and collect statistics about weekly turnover, number of employees, borrowings from banks, number of sales per day, average dollar value of a sale, ratio of cash to credit sales, and number of wholesalers with a transaction every month, we have a table of seven columns for the measures and sixty rows for the businesses. A PCA will find the most important underlying factor in clustering and separating the businesses. This might turn out to be low-price, high-sales businesses like newsagents, cafes, and corner shops versus high-price, low-sales businesses like restaurants and whitegoods retailers. Having accounted for most of the (otherwise bewildering) differences between the businesses, the process then looks for a second independent factor to help explain those differences not accounted by the first – which, in this example, might turn out to be the difference between businesses located uptown and businesses downtown. The analysis has provided a way to see some simple but strong patterns in what is initially a confusing mass of data.

In his book Burrows presents numerous striking examples of the fruitfulness of the analysis of very common words by PCA and other simpler methods. The texts, often divided into segments, are mapped on charts according to their scores on the principal components, reflecting each text

[33] Burrows, *Computation*, 3.

[34] Karl Pearson, 'On Lines and Planes of Closest Fit to Systems of Points in Space', *Philosophical Magazine* 2.6 (1901), 559–72; and Harold Hotelling, 'Analysis of a Complex of Statistical Variables into Principal Components', *Journal of Educational Psychology* 24.6 (1933), 417–41.

or segment's use of the chosen words and the weightings of those words. A PCA chart of the stages of Anne Elliot's dialogue and reported thoughts in *Persuasion* shows that the segments vary only a little in the early phases of the novel, in line with her confinement in what Mary Lascelles calls 'the prison that Sir Walter and Elizabeth have made of Kellynch'.[35] However, Burrows notes,

> As the gates of the prison begin to yield, the reader can see more room for hope than Anne has cause to do. But the movement from A4 to A6 [the fourth to the sixth of eight segments of her *spoken dialogue*] shows that she becomes free, at least, to talk more freely. For her as for Fanny, the last two phases show a more settled speech-idiolect. But her thinking changes still. Notwithstanding small fluctuations as her hopes of Wentworth rise and fall, her thought-idiolect increasingly approximates to the rhythms of speech. In a6 and a8 [the sixth and eighth of eight segments of her *reported thought*] especially, the accents of the inner voice are scarcely more stiff and formal than her speech-idiolect had been at the beginning. This is a mood of 'smiles reined in and spirits dancing in private rapture' (p. 240), a mood more exquisitely portrayed in its main lines and more fully realized in the very texture of Jane Austen's language than any of the moods that resemble it in the earlier novels.[36]

Burrows was careful to distinguish between the facts of numerical counts and the interpretations which follow. An early chapter in his book canvasses many instances of characters' use of the first-person plural pronouns, with detailed discussion of the local contexts. He notes that:

> The gulf in comparative incidence between the opposite extremities of the scale that underlies the foregoing discussion is a matter of demonstrable fact, to which we shall return. The differences between the actual pronoun idioms of the various characters lie in the more open ground of literary inference and interpretation. So far as literary interpretation is well founded, they can be seen as illuminating the 'personality' and 'situation' of each character that has been discussed. This, obviously, is not to suggest that my particular interpretations have any claim to be definitive. It is rather to insist that, even with such inconspicuous words as 'we', 'our', and 'us', worthwhile interpretative possibilities arise and that, in the further matter of literary evaluation, Jane Austen's long-standing reputation for exactitude and for 'density of texture' is given fresh support.[37]

[35] Mary Lascelles, *Jane Austen and Her Art* (Oxford: Clarendon Press, 1939), 181.
[36] Burrows, *Computation*, 211, referring to the edition of *Persuasion* in R. W. Chapman (ed.), *Jane Austen's Six Novels*, 5 vols., 3rd illus. edn (Oxford University Press, 1932–5).
[37] Burrows, *Computation*, 28.

Burrows' first experiments were with a form of PCA that started by correlating observations (businesses in our example). With words data – with frequencies declining very rapidly from the commonest, like *the* and *and*, to the less common words whose counts will be much lower – this means that the first principal component is dominant, accounting for most of the variation, and is largely a measure of how closely each text follows the average profile for word counts. A friend, the computer scientist Professor Christopher (Chris) Wallace, later suggested that Burrows should use the correlation PCA. This starts by correlating the variables – that is, it first establishes how similar the frequency-patterns for different words are to one another – and then finds the principal components of the resulting table of correlations. Each of the chosen word-variables therefore plays an equal part in the analysis, regardless of its comparative abundance or scarcity in the texts.

PCA gave Burrows access to the interactions of frequencies of very common words in the texts. Some of these appear regularly together, such as *thou* and *thee* in early modern English. If the word *thou* appears regularly in a given set of texts, then *thee* is likely to appear too. Some less obvious pairings emerge as well. *The* and *of* tend to appear together. Where texts included a lot of nouns with *the* as determiner, the preposition *of* was likely to be common as well, in texts that specify and elaborate. In a set of novels, *the* and *of* could be a useful index, arranging texts from the most descriptive, with high scores, to the ones focusing on action, and dialogue, with low scores.

The combination of PCA and very common words was the key method for what came to be known as computational stylistics. It also proved useful in separating authors. In a review of the field in the 1990s, David I. Holmes described it as the 'standard first port-of-call' in quantitative authorship attribution.[38]

Function Word Frequencies and Style

At the core of computational stylistics as Burrows developed it is the claim that frequencies of the very common words are a useful index of style. These words, which tend to be function words, have advantages operationally. They are easy to count and appear regularly, so that they give access to deep-seated steady variation. In sheer bulk they account for a good chunk

[38] David I. Holmes, 'The Evolution of Stylometry in Humanities Scholarship', *Literary and Linguistic Computing* 13.3 (1998), 114.

of all the words in a text: the 221 function word forms (or 'types') on our usual list are approximately 58 per cent of all the word tokens in a 251-play dataset.[39] Lexical words bear more obvious meaning, and are much more likely to be noticed by a reader, but they appear sporadically and as the result of contingencies like topic and setting, and so are harder to link to a persisting and large-scale style. To pursue them runs the risk of taking too much account of the accidental and local. In this vein T. S. Eliot remarked that 'Comparison and analysis are the chief tools of the critic', tools 'to be handled with care, and not employed in an inquiry into the number of times giraffes are mentioned in the English novel'.[40]

As already noted, the persistence and commonness of function words means that they normally go unnoticed. Yet changes in frequencies of these words generally signal a significant change in construction or orientation. Thus the argument sometimes made that frequency does not mean salience – that an exceptional, foregrounded use of a linguistic item may have a larger effect of a series of repetitions – has less force with function words. Burrows offers some examples from the language of Jane Austen:

> However narrow the linguistic function of words like these, it is evident that if, as is indeed the case, disparities like these are typical of the language of Jane Austen's major characters, the effects must colour every speech they make and leave some impression in the minds of her readers. Even for the most attentive novel-reader, such an impression need not – and seldom does? – consist in a definite recognition that someone is peculiarly given, for example, to the use of 'I' and 'not' and has little recourse to 'the' or 'of'. It would ordinarily consist in an awareness, however inarticulate, of the larger implications – grammatical, semantic, psychological, social – that are marked by such peculiarities. Statistical analysis of the peculiarities of incidence makes it possible to approach the whole penumbra of 'meaning' in a new and fruitful way.[41]

English is what linguists call an 'analytic language', in that its grammatical relationships are mostly conveyed by function words rather than, as in so-called 'synthetic languages', being conveyed by lexical words inflected to indicate case, gender, tense, mood, and so on. This aspect has been underlined by cognitive grammar. The schema 'container' is signified by

[39] The 221 function words are listed in Appendix E. The terms 'type' and 'token' are used to differentiate between the word as an abstract entity and the concrete, particular instances of that word. Thus understood, the phrase 'to be or not to be' contains four types (*to*, *be*, *or*, *not*) and six tokens (because there are two instances of the types *to* and *be*).

[40] T. S. Eliot, 'The Function of Criticism' (1923), in *Selected Essays*, 2nd edn (London: Faber and Faber, 1932), 32–3.

[41] Burrows, *Computation*, 4.

the function word *in* in the prepositional phrase 'in the room'.[42] Potential versus actual is signified in English by the modal verbs *may* and *might*. As easy as these so-called 'small words' are for readers to ignore, their power to shape meaning has been demonstrated in recent years by Lynne Magnusson and Sylvia Adamson.[43]

The English pronoun system encodes number (*he* versus *they*), case (*he* versus *him*), person (*he* versus *I*), and intimacy (*thou* versus *you*), not systematically, but strongly. *Her* can be possessive or objective, and *you* can be singular or plural, but otherwise there is generally a pronoun form for each of the primary syntactic categories. Propositional verbs ('give up'), past and present modal forms (*are* versus *were*), and a complex article system (*the*, *a*, *an*, *some*, zero article [i.e., omitting the article entirely], and so forth) all contribute to the set of readily countable and meaning-heavy function words. In computer-aided analysis, the 'tagging' or 'marking up' of homograph forms – such as *to* as in 'to Sydney' and *to* as in 'to act', *her* as in 'her hand' and as in 'she promised her', and the three forms of *that* in 'he was able to tell her that that key was the one that opened the second box' – to formalise these distinctions further enhances the machine-readability of function words.

It is not immediately obvious that the frequencies of function words – simply counting occurrences, taking no account of the sequences in which they are placed – could serve to differentiate styles. Yet because of their structural roles in sentences they do bear traces of patterns of construction. Franco Moretti, for instance, found that titles of anti-Jacobin novels began with definite articles – *The Democrat*, *The Infidel Father* – far more often than those of New Woman novels, which favoured titles beginning with indefinite articles – *A Bluestocking*, *A Hard Woman*, and so on. Moretti argues that this can be explained by the different structural uses of the articles. The definite article is used with a known entity, fitting the anti-Jacobin defence of the status quo; the indefinite article, which typically introduces something new, serves the New Woman agenda of support for change.

> So: *A Girton Girl*, *A Hard Woman*, *A Mummer's Wife*, *A Domestic Experiment*, *A Daughter of Today*, *A Semi-detached Marriage*: what the article says

[42] Louisa Connors, 'Computational Stylistics, Cognitive Grammar, and *The Tragedy of Mariam: Combining Formal and Contextual Approaches in a Computational Study of Early Modern Tragedy*', Ph.D. thesis (University of Newcastle, 2013), 86–90.

[43] Sylvia Adamson, 'Understanding Shakespeare's Grammar: Studies in Small Words', in Sylvia Adamson, Lynette Hunter, Lynne Magnusson, Anne Thompson and Katie Wales (eds.), *Reading Shakespeare's Dramatic Language: A Guide*, (London: Arden Shakespeare, 2001), 210–36; and Lynne Magnusson, 'A Play of Modals: Grammar and Potential Action in Early Shakespeare', *Shakespeare Survey* 62 (2009), 69–80.

> is that we are encountering all these figures *for the first time*; we think we know what daughters and wives are, but we actually don't, and must understand them afresh. The article announces the novel as a challenge to received knowledge. And instead, the democrat, the Parisian, the infidel father . . . We know these people! Anti-Jacobin titles don't want to change received ideas, they want to *use* them: the French Revolution has multiplied your enemies – beware.[44]

MacDonald P. Jackson provides another example of the way the frequencies of function words reflect different styles:

> Consider two differently constructed sentences that convey the same information. Here is the first: 'As soon as we guests had finished dinner, we said goodbye to our kind hosts and drove to the theatre, where we saw a performance of *Twelfth Night*, which we greatly enjoyed.' And here is the second: 'Straight after dinner, we guests, saying goodbye to our kind hosts, drove to the theatre and saw a most enjoyable performance of *Twelfth Night*.' The two sentences each contain two examples of 'to' and one of 'a', 'and', 'of', 'our', and 'the'. But the first has three more instances of 'we' than the second, and also contains 'as' (twice), 'had', 'where', and 'which', none of which are found in the second sentence, which has instances of 'after' and 'most', both absent from the first. The two types of sentence construction entail the use of different function words. The first sentence uses a relative clause, introduced by 'which', whereas the second does not. The first sentence uses the conjunction 'and' to link co-ordinate clauses 'we said . . . and drove', whereas the second modifies 'we guests' by using the present participle 'saying'.[45]

The gist of Jackson's two invented sentences is the same, but they are constructed differently, and we can hazard a stylistic interpretation. The first sentence is more pedantic and more focalised through *we*, while the second follows a less predictable sequence and moves more decisively to the performance of the play as its destination. Jackson tallies the presence and absence of the various function words in the two constructions and their frequencies to illustrate the basis for the functioning of a typical computational-stylistics analysis, which happens in reverse: there we start with the patterns in the function word frequencies, and infer from them something of the style of the samples.

Quantitative Work on Style in Early Modern English Drama

Whatever the appeal of this and other computational-stylistic methods, the number of quantitative stylistic studies within the field of the present

[44] Franco Moretti, 'Style, inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850)', *Critical Inquiry* 36.1 (2009), 154–6.

[45] MacDonald P. Jackson, *Determining the Shakespeare Canon: 'Arden of Faversham' and 'A Lover's Complaint'* (Oxford University Press, 2014), 42.

book – that is, early modern English drama – has been small, if strictly author-attribution work is excluded. A number of scholars have studied statistical patterns in dramatic verse.[46] Dolores M. Burton compared the grammatical styles of two Shakespeare plays.[47] Some individual function words have been studied in Shakespeare and elsewhere.[48] Two studies have challenged the traditional belief that Shakespeare's vocabulary was exceptionally large.[49] There is a corpus-based study of the rhetoric of *Hamlet*,[50] as well as studies of the interactions in the first scene of *King Lear* and of phrasal repetends in *Troilus and Cressida*.[51] Staying with Shakespeare, but ranging more widely across the canon, scholars have applied computational methods to analyse genre in Shakespeare,[52] to his late style,[53] to his characterisation,[54] and to the varying length of speeches in the plays.[55] There is quantitative work on the distribution of props in Shakespeare and

[46] Representative examples include Philip W. Timberlake, *The Feminine Ending in English Blank Verse* (Menasha: George Banta, 1931); Ants Oras, *Pause Patterns in Elizabethan and Jacobean Drama: An Experiment in Prosody* (Gainesville: University of Florida Press, 1960); Marina Tarlinskaja, *Shakespeare and the Versification of English Drama, 1561–1642* (Farnham: Ashgate, 2014); and Douglas Bruster and Geneviève Smith, 'A New Chronology for Shakespeare's Plays', *Digital Scholarship in the Humanities* 31.2 (2016), 301–20.

[47] Dolores M. Burton, *Shakespeare's Grammatical Style: A Computer-Assisted Analysis of 'Richard II' and 'Antony and Cleopatra'* (Austin: University of Texas Press, 1973).

[48] For example, Ulrich Busse, *Linguistic Variation in the Shakespeare Corpus: Morpho-Syntactic Variability of Second Person Pronouns* (Amsterdam: John Benjamins, 2002); and Hugh Craig, 'Plural Pronouns in Roman Plays by Shakespeare and Jonson', *Literary and Linguistic Computing* 6 (1991), 180–6, 'Grammatical Modality in English Plays from the 1580s to the 1640s', *English Literary Renaissance* 30.1 (2000), 32–54, and '*A* and *an* in English Plays, 1580–1639', *Texas Studies in Literature and Language* 53.3 (2011), 273–93.

[49] Hugh Craig, 'Shakespeare's Vocabulary: Myth and Reality', *Shakespeare Quarterly* 62.1 (2011), 53–74; Ward E. Y. Elliott and Robert J. Valenza, 'Shakespeare's Vocabulary: Did It Dwarf All Others?' in Mireille Ravassat and Jonathan Culpeper (eds.), *Stylistics and Shakespeare's Language: Transdisciplinary Approaches* (London: Continuum, 2011), 34–57.

[50] Thomas Anderson and Scott Crossley, '"Rue with a Difference": A Computational Stylistic Analysis of the Rhetoric of Suicide in *Hamlet*', in Mireille Ravassat and Jonathan Culpeper (eds.), *Stylistics and Shakespeare's Language: Transdisciplinary Approaches* (London: Continuum, 2011), 192–214.

[51] Dawn Archer and Derek Bousfield, '"See Better, Lear?" See Lear Better! A Corpus-Based Pragma-Stylistic Investigation of Shakespeare's *King Lear*', in Dan McIntyre and Beatrix Busse (eds.), *Language and Style* (Basingstoke: Palgrave, 2010), 183–203; Ian Lancashire, 'Probing Shakespeare's Idiolect in *Troilus and Cressida*, 1.3.1–29', *University of Toronto Quarterly* 68 (1999), 728–67.

[52] Hope and Witmore, 'The Hundredth Psalm'.

[53] Michael Witmore and Jonathan Hope, 'Shakespeare by the Numbers: On the Linguistic Texture of the Late Plays', in Subha Mukherji and Raphael Lyne (eds.), *Early Modern Tragicomedy* (Woodbridge: Boydell and Brewer, 2007), 133–53.

[54] Hugh Craig, '"Speak, That I May See Thee": Shakespeare Characters and Common Words', *Shakespeare Survey* 61 (2008), 281–8; Jonathan Culpeper, 'Keywords and Characterization: An Analysis of Six Characters in *Romeo and Juliet*', in David L. Hoover, Jonathan Culpeper, and Kieran O'Halloran (eds.), *Digital Literary Studies: Corpus Approaches to Poetry, Prose and Drama* (New York: Routledge, 2014), 9–34.

[55] Hartmut Ilsemann, 'More Statistical Observations on Speech-Lengths in Shakespeare's Plays', *Literary and Linguistic Computing* 23.4 (2008), 397–407.

the early modern drama,[56] and one attempt to apply computational methods to problems in the transmission of play-texts.[57]

Challenges to Stylistics

One fundamental objection to stylistics is based on the conviction that literature is always more than the sum of its constituent parts – that numerical methods, which are inevitably reductionist, can offer nothing useful for literary analysis.[58] Yet style does have a numerical aspect, and stylistics is founded on this truth. Each time a word is used, its meaning is created afresh, and is thus unique, yet it is also recognisably an instance of that word, a token of that word type. The quantitative analysis of style depends by definition on defining a language feature – at the simplest level, a word type – and then counting instances of that feature as if they were all the same. When the features are being used for a classifier this practice is easy to defend. If the feature-counts in whatever combination do serve to separate known members of the classes introduced as test samples, then the procedure is validated and scepticism about the categories is quietened.

Each time Prospero addresses Miranda, Caliban, Ariel, and Fernando as 'thou', this pronoun has a peculiar inflection more or less consciously apparent to audiences and readers, but it is also a choice within a system of pronoun types – most immediately, a choice between 'thou' and 'you' forms[59] – and therefore susceptible of a wider analysis in the context of other characters, plays, and canons. This is not the realm of classification but of a continuously varying spectrum of frequency, which can take its part in a network of variation with other words and other language features. How much such patterns illuminate literary questions is always a matter of debate. Language provides a rich source for statistics, as words are repeated or not, appear often or rarely near each other, and so on, but this abundance is no guarantee of interest. Independent of the methods employed, it is up to the literary critic to 'know when to compare and when to analyze',[60]

[56] Frances Teague, *Shakespeare's Speaking Properties* (Cranbury: Associated University Presses, 1991); Douglas Bruster, 'The Dramatic Life of Objects in the Early Modern English Theater', in Jonathan Gil Harris and Natasha Korda (eds.), *Staged Properties in Early Modern English Drama* (Cambridge University Press, 2002), 67–96.

[57] Lene B. Petersen, *Shakespeare's Errant Texts: Textual Form and Linguistic Style in Shakespearean 'Bad' Quartos and Co-Authored Plays* (Cambridge University Press, 2010).

[58] See de Man, 'Literary History', and Willie van Peer, 'Quantitative Studies of Literature: A Critique and an Outlook', *Computers and the Humanities* 23.4–5 (1989), 301–7.

[59] See Roger Brown and Albert Gilman, 'The Pronouns of Power and Solidarity', in Thomas A. Sebeok (ed.), *Style in Language* (Cambridge, MA: MIT Press, 1960), 253–76.

[60] Eliot, 'The Function of Criticism', 33.

guided by the accumulated understanding of the discipline in determining which among the plethora of possibilities is worthy of attention.

Another version of this objection is that features in literary study cannot be counted because meaning is constructed by the reader: instances of a given word on a page may appear to be stable or commensurate, the argument goes, but in fact they are 'relational', and counting them as if they were equivalent is misleading.[61] This is an important objection to consider. If sustained, it would invalidate all quantitative stylistics at a stroke. Indeed, Stanley Fish has repeatedly denounced stylistics in these terms since the 1970s.[62] As a blanket objection it is probably not sustainable, since to do so would rule out statistics in general – that is, any attempt to encapsulate events through counting. It is always possible to see individual variations in the instances accumulated, but there are important benefits in grouping them together. To sustain the objection, one would have to show that literary data is somehow intrinsically impossible to put into categories. This is, in essence, Fish's approach. One might respond that some features of literary data can be classified in categories, and some cannot. It is easy to think of cases in literary study where counting could be done in invalid categories. For example, counting instances of words with unrelated senses like *spring* or *lead*, or counting cases where a character waves a hand or any other action for which there are multiple and differing contexts and intentions. Since a sword and a pen have different associations in the literatures of various periods, counting instances and comparing the numbers indiscriminately would be of limited worth. On the other hand, one might defend counting instances of the word *all* on the grounds that each shares enough of a semantics to make them commensurate.

Burrows had already tackled this question in *Computation into Criticism*. He considers the objection that words 'mean nothing' if taken out of context:

> In answer to such an objection, a traditional grammarian would maintain that 'we', like other words, should be regarded as having incipient meaning, in a sort of Aristotelian potentiality, not realized until it is set in context. More recent authorities, following Roman Jakobson, would maintain that, for any speaker of English, 'we' bears a certain 'context' even before

[61] John Frow, *Genre* (New York: Routledge, 2006), 125. See also Tony Bennett, 'Counting and Seeing the Social Action of Literary Form: Franco Moretti and the Sociology of Literature', *Cultural Sociology* 3.2 (2009), 277–97.

[62] Fish, 'What Is Stylistics'; and, 'What Is Stylistics and Why Are They Saying Such Terrible Things about It? Part II', in *Is There a Text in This Class?* (Cambridge, MA: Harvard University Press, 1980), 246–67.

> it is brought into use. It is among those words that can open a sentence. It is among those words that can stand as subject to a verb. It is among those words that allude to more than one referent (the speaker being among them) without actually naming those referents. Already it is distinct from 'John', 'I', 'you', and 'us' – to say nothing of 'although' and 'purple': for none of these words conforms even to this rudimentary set of constraints on meaning . . . As soon as it is mentioned, even if it is the opening word of a fresh discourse, 'we' takes on a more immediate meaning by identifying its referents: the bases of identification, not always unambiguous at first, are likely to be predominantly grammatical when other utterances have led on to the 'we', predominantly social when it initiates a fresh discourse. On either traditional or more recent doctrine, 'we', taken alone, is not devoid of meaning.[63]

Literary language, together with the paratextual materials of literary works, provides a wide range of features to count, and thus choices must be made. The choice must then influence results, and critics of quantitative work have argued that this undermines any claims to objectivity.[64] This is a fundamental critique of *any* quantitative study beyond the hard sciences, whether it is in ecology, sociology, or psychology. It is easy to show that there are cases where the choice of features does not determine the results, so there can be surprises and a definite gain in knowledge. If we were to ask, for instance, 'Do women write differently from men?', we have a way of validating the choice of features. If the pattern of use of a given feature shows a significant difference in balanced and commensurate sets of samples of the writing of women and the writing of men, then it does not matter how the unit was chosen. Here we have an external basis on which to discard some features and accept others: the difference between two objectively based classes.

There is a claim that Shakespeare's later verse is more informal and conversational than his earlier efforts. To investigate this claim, we might choose a group of units that intuitively seem to mark informality in verse (e.g. enjambment, contractions, hypermetric syllables, and second-person pronouns). We could then check this intuition by counting these features in groups of samples that are by consensus formal or informal. If we found that the first three are markedly and consistently more common in informal samples, but the last occurs about as often in both, we could count instances of the first three in early and late Shakespeare. The final stage would be to combine the three scores mathematically, say, by

[63] Burrows, *Computation*, 28–9.
[64] See, for example, Bennett, 'Counting and Seeing', esp. 290–1.

adding them together, to give a single value. If this separated later Shakespeare segments from earlier ones, we could show the extent of the difference between the two sets, as well as confirm that there is a genuine difference.

We can also proceed inductively. We look at all the word types used in *Jane Eyre* and a rewriting of this novel for late twentieth-century young adult audiences and determine which are used at significantly different rates between the two novels. The original set of features – the word types – has been our choice, but the selection within them is directed by patterns in the data. Then there are cases of classification, such as by author and by date. We can seek markers of the classes, check them with known members of the classes, and then find the patterns for the chosen markers in disputed cases. We have an objective way of validating the choice of features, so we do not care much about where they came from.

Computational stylistics generally counts the frequencies of particular words, which are as close as possible to a 'given'. Nevertheless, other linguistic features could be counted: the letters of which words are composed, for example, or punctuation; combinations of words; marked-up features, such as images and figures of speech, and so on.

Given 'world enough and time', quantitative analysis could perhaps proceed without having to select features to count and discard others. Even so, some features are not susceptible to counting, and thus the results can be characterised as representative but not universal. Stylistics captures significant aspects of style, but not the totality. Any findings are relative to the features chosen rather than absolute. In computational stylistics, practitioners sometimes speak of a style, referring to collective patterns of particular features. It is important to remember that literary style in the general sense encompasses so much more than that. If we make a profile of the usage of the 100 most common words, it may be revealing, but only in specific ways, and it certainly does not capture everything that might be included under 'style'.

The Variability and Predictability of Literary Language

Commentators on literary language, as on all language, may choose to focus either on the variability of language or its predictability. There is a premium on creativity in language. Even in commonplace exchanges we expect variation, as a guarantee of spontaneity and full and conscious investment in the present. Readers of literary texts expect to be surprised occasionally, even if against a background of probability.

Yet language is also a shared knowledge, so variation must be limited if it is to function as communication. Beyond that, economy of effort demands that most of what we read or hear must be already familiar. As soon as we recognise that a character is a Petrarchan lover, a large set of associations is invoked and can be used to create effects of familiarity or reversal. We are listening to a friend tell a story about what happened yesterday, but we are also listening for the familiar idiosyncratic characteristics of the speaker to be rehearsed and responding to a well-established framework of expectations which make the story work through surprise, amusement, or sympathy.

The language of literary texts is endlessly creative but also manifests some regularities, so that it is predictable in relation to some categories such as author, genre, period, mode, and so on. Each time an individual writer at a certain moment takes up a pen to write in a certain familiar genre, he or she is free to put old or new ideas in old or new ways, but also cannot help writing as that person at that time in that genre. It is impossible to predict with absolute precision what the resulting writing will be, but it is possible to formulate some ranges beyond which it is unlikely to go, and retrospectively, given an already written sample of unknown author, genre, date, and so on, these constraints allow the observer to place the sample with a good degree of reliability within some broad categories.

This balance of predictability and unstructured fluctuation puts literary language into the category of a 'stochastic' system, from the Greek στόχος, 'aim'. Some broad directions or regularities are evident, while each component step cannot be precisely predicted. The content of a sentence spoken in a play is certainly not precisely predictable – language is not a deterministic system – but those contents do take part in a pattern of regularity that constrains variation. The same can be said of many aspects of human activity. No one murder is entirely predictable, and citizens exercise free will, but the number of murders in a given large city in a year can be predicted with impressive accuracy.[65]

Immanuel Kant had already grasped this fundamental insight of statistics in the eighteenth century. In his 'Idea for a Universal History' (1784), Kant allies it to a universal human nature – which we might cavil with – but the perception that a change of scale may reveal supervening patterns where close observation reveals only unfettered individual choice remains applicable:

> we know that history, simply by taking its station at a distance and contemplating the agency of the human will upon a large scale, aims at unfolding

[65] Ian Hacking, *The Taming of Chance* (Cambridge University Press, 1990), 41.

> to our view a regular stream of tendency in the great succession of events; so that the very same course of incidents, which taken separately and individually would have seemed perplexed, incoherent, and lawless, yet viewed in their connexion and as the actions of the human *species* and not of independent beings, never fail to discover a steady and continuous though slow development of certain great predispositions in our nature. Thus for instance deaths, births, and marriages, considering how much they are separately dependent on the freedom of the human will, should seem to be subject to no law according to which any calculation could be made beforehand of their amount: and yet the yearly registers of these events in great countries prove that they go on with as much conformity to the laws of nature as the oscillations of the weather: these again are events which in detail are so far irregular that we cannot predict them individually; and yet taken as a whole series we find that they never fail to support the growth of plants – the currents of rivers – and other arrangements of nature in a uniform and uninterrupted course.[66]

Here the large-scale flow of history makes a neat parallel with Puttenham's vision of style as a quality only apparent over the full sweep of a literary work. Probability emerged as the key concept in the unfolding of this principle of apparently freely fluctuating local events and larger regularities in the decades following Kant's observation, and came to dominate physics as well as sociology by the first half of the twentieth century. Newton's laws of motion, articulated in the late seventeenth century, were strictly causal and had no need of probability: action determined reaction and could be predicted by an equation. By contrast, James Clerk Maxwell's nineteenth-century observations on the motion of molecules depended on patterns and accumulations: no individual movement is predictable, but the combined effect is (within a limited range of fluctuation).

It was not until well into the nineteenth century that measurement was embraced as a core aspect of the natural sciences, and this paved the way for the view that in the natural and human worlds 'laws of chance' provided a better explanation for events than 'strictly causal laws' – that is, 'equations with constant numbers in them'.[67] The fundamental discovery of computational stylistics as developed by John Burrows is that literary language, too, is stochastic.[68]

[66] Immanuel Kant, 'Idee zu einer allgemeinen Geschichte in weltbürgerlicher Absicht', trans. Thomas de Quincy, 'Idea of a Universal History on a Cosmo-Political Plan', *The London Magazine* Oct. 1824, 385–93 (385).

[67] Hacking, *Taming of Chance*, 5, 1, 63.

[68] Willard McCarty, 'Getting There from Here: Remembering the Future of Digital Humanities', *Literary and Linguistic Computing* 29.3 (2014), 283–306, esp. 289.

Language is a creative human production, and from one perspective a corpus – such as a collection of the dialogue of early modern plays – is best seen as flux, wildly gyrating cross-currents, the competing, endlessly inventive voices of thousands of characters from tavern hostesses to duchesses and from base villains to stout heroes. Yet there are patterns that emerge from the flux, regularities in categories of plays and scenes and characters. Along with probabilistic thinking, philosophical pragmatism is one way to conceptualise this relationship of chaos and pattern. In Lars Engle's words, pragmatism suggests that 'strata of stable contingency underlie and shape the liquid flow of experience and the volatility of thought'.[69] Frank Lloyd Wright's Imperial Hotel in Tokyo was built on a subsoil of mud in an earthquake-prone location, but achieved stability by having the building rest on concrete rafts floating in the mud and allowing independent movement.[70] In the same way, numerical analysis allows us to see that in the flux of language there are continuities and predictable patterns, by no means absolute, but resiliently present nevertheless.

The statistical findings in this book are themselves a challenging mixture of certainty and ambiguity. At the level of numbers there is certainty, with some caveats. For example, in the first 1,000 running words of *Hamlet* there is an exact number of instances of the word *the*, provided we specify one particular version of the play and some rules, such as that 'th'end' contractions are regularised to 'the end' and 'th'art' contractions are regularised to 'thou art', so that one yields an instance of *the* and the other does not.

At a second level of processing we might compare the counts for *the* or some other word for 10 Shakespeare plays with counts in 100 plays we know are not by Shakespeare, in each case converting the counts to a figure for how many instances per 100 words. Perhaps the average Shakespeare score is higher than the average score for the others. We can say with certainty (as long as the caveats about text and modernisation are borne in mind) that this is so, but the question then follows, *is this difference significant*? Are the Shakespeare scores consistently higher than non-Shakespeare ones, or is it more a matter of the occasional extreme raising the average? Do scores fluctuate wildly as a result of local variation in theme or style, so that we cannot rely much on differences across a canon, or are they relatively steady, so that a consistent high or low score demands interpretation? Here statistics can help with measures of difference between two groups of scores, taking into account the degree of difference and the degree of scatter.

[69] Lars Engle, *Shakespearean Pragmatism: Market of His Time* (University of Chicago Press, 1993), 37.
[70] Engle, *Shakespearean Pragmatism*, 233 n.27.

Nevertheless, at this point we are well and truly in the domain of relativity and judgement. The statistics can only give a probabilistic finding by telling us that in the universe of similar patterns, this difference will only rarely come about as a chance effect rather than a true underlying contrast. We get an estimate along a spectrum rather than an absolute 'yes' or 'no'. Then, even when there is a very large difference with a small chance that this is simply random variation, there are more questions: this may be statistically highly significant, but would it mean anything to readers, and how could we explain it in terms of the writer? What about other Shakespeare plays, and other plays by other writers? If we wish to think of this as an immutable Shakespeare characteristic, what about his published poems, and what about the plays and poems he wrote which did not survive, and what about the works he was capable of writing but did not?

This Book

In this book, we offer a series of largely independent treatments of some specific literary-historical questions. After detailing our methods in Chapter 1, we assess how far the medium of verse itself governs style, both in all-verse plays and in plays that mix verse and prose in Chapter 2. Chapter 3 analyses the plays by character, highlighting characters from plays by different authors whose dialogue styles are very similar, suggesting that they occupy the same dramatic niche. In Chapter 4, we move away from dialogue to look at the distribution of props in plays staged in professional theatres between 1590 and 1609. Do authorship and genre have an effect on the use of props? Chapter 5 focuses on chronology and highlights collective change in dramatic dialogue from the 1580s to the 1630s. Chapter 6, like Chapter 2, examines a long-standing belief about broad patterns in style in early modern drama – the claim that repertory companies cultivated a distinctive style, analogous to an authorial style. The final chapter, Chapter 7, moves beyond the immediate period to examine how comedies and tragicomedies of the 1660s compare stylistically with their pre-Restoration counterparts. As a coda, we consider the implications of the findings as a group and sketch promising avenues for future work, and appendices detail the plays, characters, prop-lists, and function words we have used.