'SPEAK, THAT I MAY SEE THEE': SHAKESPEARE CHARACTERS AND COMMON WORDS

HUGH CRAIG

Recording the number of times a particular word or phrase is used in a passage, or the relative frequency of a metrical feature or of a rhetorical figure, has been a familiar practice in Shakespeare studies for at least a century. This sort of measurement and comparison is a staple of authorship studies. Along the way these practices have suffered some swingeing blows, such as Sir E. K. Chambers's attack of 1924 on the 'disintegrators',[1] Samuel Schoenbaum's ridicule of the 'parallelographic school' in the 1960s,[2] and Brian Vickers's recent demolition of the case for Shakespeare's authorship of *A Funerall Elegie*.[3] Nevertheless, they have proved indispensable (when soundly practised) as a complement to documentary evidence and to subjective estimates of what is, or is not, the authentic style of a particular writer. Statistical work plays a large part in Gary Taylor's *Textual Companion* to the Oxford Shakespeare (Oxford, 1980), and in important recent books by Vickers himself and by MacDonald P. Jackson.[4] It is sympathetically assessed in Harold Love's *Attributing Authorship: An Introduction* (Cambridge, 2002).

Much less use has been made of these methods in describing Shakespeare's style, or that of his peers, more generally, even in the age of effortless counting and calculation by the computer. Two exceptions are Barron Brainerd's work on pronouns and other common words in relation to genre and period in Shakespeare's works,[5] and Jonathan Hope's book on the sociolinguistics of Shakespeare's idiolect in contrast with Fletcher's.[6] (Hope and Michael Whitmore have recently presented the results from a more ambitious computational-stylistics venture using a complete set of Shakespeare plays.[7]) Meanwhile there are promising precedents in other areas of English studies. John Burrows's book *Computation into Criticism: A Study of Jane Austen and an Experiment in Method* (Oxford, 1987) offers remarkable insights into the patterning of Jane Austen's language, and the subtly varied idiolects of her characters. Franco Moretti has applied quantitative analysis to a wide sweep of literary history, with fascinating results, in his *Graphs, Maps, Trees: Abstract Models for a Literary History* (London, 2005). David L. Hoover has demonstrated the extraordinary consistency of the progressive changes in the style of Henry James's novels through computational stylistics.[8]

---

[1] *The Disintegration of Shakespeare* (London, 1924).

[2] *Internal Evidence and Elizabethan Dramatic Authorship: An Essay in Literary History and Method* (London, 1966).

[3] *'Counterfeiting Shakespeare': Evidence, Authorship, and John Ford's* Funerall Elegie (Cambridge, 2002).

[4] Vickers, *Shakespeare, Co-Author: A Historical Study of the Five Collaborative Plays* (Oxford, 2003) and *Shakespeare,* A Lover's Complaint, *and John Davies of Hereford* (Cambridge, 2007); Jackson, *Defining Shakespeare:* Pericles *as Test Case* (Oxford, 2003).

[5] 'The Chronology of Shakespeare's Plays: A Statistical Study', *Computers and the Humanities* 14 (1980), 221–30, and 'Pronouns and Genre in Shakespeare's Drama', *Computers and the Humanities* 13 (1990), 3–16.

[6] *The Authorship of Shakespeare's Plays: A Sociolinguistic Study* (Cambridge, 1994).

[7] 'The Very Large Textual Object: A Prosthetic Reading of Shakespeare', *Early Modern Literary Studies,* 9 (2004), 6.1–36 (http://purl.oclc.org/emls/09-3/hopewhit.htm).

[8] 'Corpus stylistics, stylometry, and the styles of Henry James', *Style* 41 (2007), 160–89.

Part of the hesitation of Shakespearians in adopting the new methods must arise from doubts about the validity and value of assessing dramatic language like Shakespeare's through computation, and from uncertainty about how to interpret the constructs which the calculations produce. Can a statistical method be properly sympathetic to the richness and subtlety of literary language? Assuming that the data collection, and the procedures, are sound, can the results escape banality, and can any commentary on them be anything other than tautology, or wild speculation? While most would accept the usefulness of statistics in epidemiology, or market research, its application to literature still seems a barbarous practice. Stanley Fish's two articles on stylistics deny that the quantitative study of style can have any usefulness at all.[9] Even more sympathetic scholars like Willie van Peer have argued that the trade-off between what is countable, and what is of interest to literary scholars, must always result in a loss of all specifically literary aspects of textuality.[10] For a humanist like George Steiner computational techniques compromise the ineffable essence of literary works.[11] The post-structuralist, on the other hand, might see literary statistics as a display of extreme bad faith, insofar as it claims an application of the patterns of language to anything outside itself, and one free of ideology at that.[12]

My own view is that the computer has indeed made a difference. Given the abundance of machine-readable text now available, and the speed of processing now possible, literary statistics finally does have some particular things to offer which we can get in no other way. There is a bargain to be made. To count something, and thus secure the data on which the procedures can work, one must apportion linguistic features to a finite number of discrete categories. To do this, a thousand subtle distinctions obvious to every reader have to be ignored; but having done this, the computer can deploy a superhuman capacity to remember and to process systematically. The computer can make a representation of the textual world which is nothing like an interpretation, but is certainly directly and objectively related to that world. It offers a scientific instrument, as it were: a spectrometer, say, which could never replace human vision in understanding an object, but can yield information not available to the naked eye.

Computation is in fact in sympathy with some aspects of language. Language is inherently repetitive and works by variation against a pattern of predictability. In its written form, at least, it works as an assembly of base-level items, words, which are necessarily limited in number because they must be shared by writer and reader. A finite set of items is repeated in different combinations to produce meaning. In this sense a language like English is not only susceptible to counting – each string of characters between spaces or punctuation is a recognizable item – but works in part by sheer frequency. An abundance of the words *I* and *me* relative to the established pattern for that kind of discourse conveys important information to the hearer; consistently writing *upon* where the reader expects *on* does the same.

The present article describes an experiment in which a group of Shakespeare characters are compared on the basis of their recourse to a small group of common words. At its heart is a statistical process drawing out some key patterns in contrast and likeness in the language of the characters. These calculations are performed though an open access website at the Centre for Literary and Linguistic Computing at Newcastle (Australia) and so can be replicated, or re-run with different stipulations, by anyone with an internet connection. A reader

---

[9] 'What Is Stylistics and Why Are They Saying Such Terrible Things About It?', *Approaches to Poetics: Selected Papers from the English Institute*, ed. Seymour Chatman (New York, 1973), pp. 109–52, and 'What Is Stylistics and Why Are They Saying Such Terrible Things About It? Part ii', *Boundary 2*, 8.1 (1979), 129–45.

[10] 'Quantitative Studies of Literature: A Critique and an Outlook', *Computers and the Humanities*, 23 (1989), 301–7.

[11] *Real Presences: Is There Anything in What We Say?* (London, 1989), pp. 82–3.

[12] No critique of this kind has in fact been mounted. Thomas Merriam does relate some of the postulates of post-structuralism to computational methods in 'Linguistic Computing in the Shadow of Postmodernism', *Literary and Linguistic Computing* 17 (2002), 181–92.

can thus check the results by reproducing them, or test some of the conclusions by running a similar experiment with (say) whole texts instead of characters. The website also allows users to launch an experiment of their own with a quite different set of words, and directed to a different aspect, like chronology, instead of character, genre or gender.[13]

The study begins with the commonest words, and the dialogue of the characters who speak most words overall. In this way the predilections of the researcher are allowed minimum play, since a simple principle of frequency is followed in selections. The commonest words tend to be function words like the pronouns, articles and conjunctions, and this has other advantages. It would seem on the face of it that they take less of their colour from their context, and thus are better candidates for counting, than lexical words. One instance of *you* is interchangeable with another, whereas (one might argue) one instance of *blood* is not. Examining only characters with larger spoken parts has the advantage that local variations, arising from particular settings or situations, tend to be ironed out by a balance with a variety of other settings and situations. Larger samples like these, we can hope, allow core tendencies, which might otherwise be masked, to be revealed. Smaller characters, though of course interesting in their own different ways, will tend to the idiosyncratic; if included, their extreme departures from the norm would tend to overwhelm the steadier patterning of the larger characters.

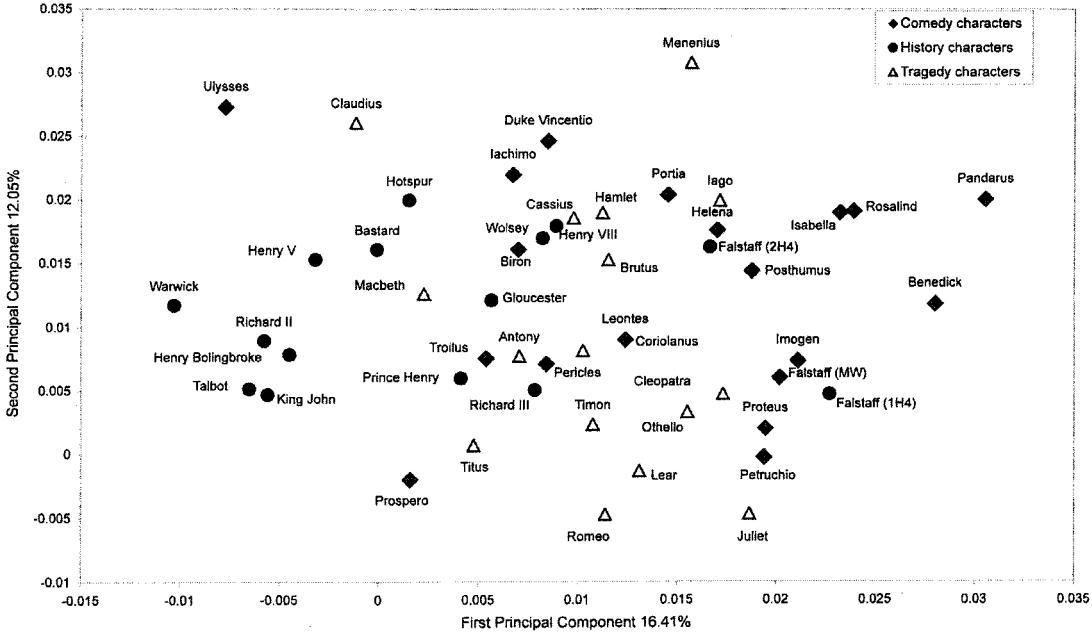
I have chosen as samples characters who speak more than 3,000 words in all (there are fifty of these), and as variables the fifty commonest words in this collection of dialogue. At the beginning, then, is a table fifty columns wide, one for each character, and fifty rows deep, one for each word. In each of the 2,500 cells of the table is a count for that word in that character's dialogue. The first step is to divide the counts by the total number of words for the character, so that (say) Hamlet's use of *you* is expressed as a fraction of his eleven and a half thousand spoken words, and can be compared with Isabella's, once her count has been divided by her total of just over 3,000 words.

To make some sense of the table we can call on a statistical procedure named Principal Component Analysis. This works to find a line of differentiation through the counts which accounts for the greatest amount of variation in it, as a process of 'data reduction'. It is as if, in a two-variable situation, one had counts for people's weights and counts for their height, and could simplify this to a single new variable, 'size', combining the two original ones. PCA looks for the new composite variable which accounts for most of the variation, then a second which accounts for the second largest amount, and so on. The new variables combine the counts from all the original variables, giving each of these contributory variables a different weighting. In a slightly more complicated case, one might take a series of counts of daily average temperature, rainfall, humidity and barometric pressure. The first principal component that emerged from this, the factor which accounts best for the various individual fluctuations of the measures, might well be related to the contrast between summer and winter, and the second to the difference between maritime and continental locations. In the case of the Shakespeare characters and the common-words data, one might expect a difference between comedy and tragedy, or perhaps early and late dates of composition, to be the strongest lines of difference. What we get is shown in illustration 23.

Here the horizontal axis is the first Principal Component, a mathematically derived 'factor' which accounts for sixteen per cent of all the variation in the table.[14] (If there were no patterns in the table, if all the variables fluctuated independently

[13] *PCA Online: The Shakespeare Computational Stylistics Facility*, www.newcastle.edu.au/cllc/pcaonline.

[14] The illustrations to the article use results from the Newcastle *PCA Online* website. This is also the source for the statistics for word-variables given here. *PCA Online* draws on the Moby Shakespeare text, a derivative of the Globe edition of the 1860s, obviously not all one would hope for in a Shakespeare text, but unambiguously in the public domain and so suitable for use in an open-access website like *PCA Online*. The Moby Shakespeare excludes *The Two Noble Kinsmen*; there would, of course, be arguments for including parts of this in a complete Shakespeare, and for excluding all



23 PCA plot of the fifty largest Shakespeare characters based on the frequency of the fifty commonest words.

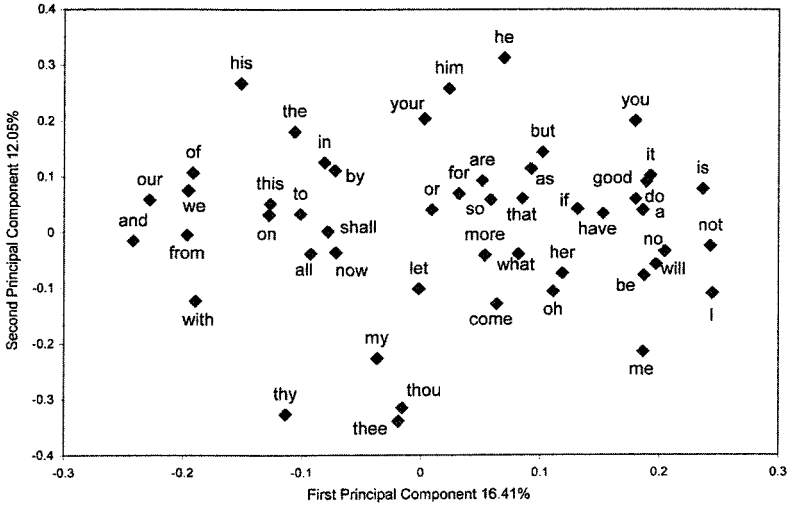
of each other, we would expect a figure of two per cent for each Component.) The vertical axis is the second Principal Component, accounting for twelve per cent.

The graph maps characters according to their use of the fifty words, weighted so as to find two lines of best fit. Pandarus from *Troilus and Cressida* and Warwick from *The True Tragedy of Richard Duke of York* are the extremes along the continuum of the First Component, the horizontal axis. We can check a cognate map of the words to see which words are most significant in forming the continuum (illustration 24).

These are the same components, this time expressed in terms of the weightings of the word-variables. At the left-hand or Warwick end the words are *and*, *our*, *we*, *from*, *with* and *of*. At the right-hand, Pandarus end the words are *I*, *not* and *is*. First-person plural pronouns are opposed to first-person singular ones. Markers of complexity and precision in description and argument are set against markers of contradiction and immediacy. In Douglas Biber's study of a range of modern writing and speech, he found that 'high informational density and exact informational content versus affective, interactional, and generalized content' was the primary factor of differentiation.[15] This contrast shares characteristics with the First Principal Component in this characters study, which pits markers of informational density (to the left-hand end of illustration 24) against words common in intensive interactions (to the right-hand end). It may be that this contrast between disquisitory and dialogic styles is a general feature which emerges in most large mixed language samples, rather than a peculiarly Shakespearian one. The words suggest impersonal, collective authority to the left, and individual assertion and contradiction to the right.

or parts of a number of plays that are represented, such as the *Henry VI* plays, *Macbeth* and *Pericles*.

[15] *Variation across Speech and Writing* (Cambridge, 1988), p. 107.



24 PCA plot of the fifty commonest words in the fifty largest Shakespeare characters.

Pandarus (to the right in illustration 23) is a character formed of negation, querulously undercutting and anxiously re-directing.

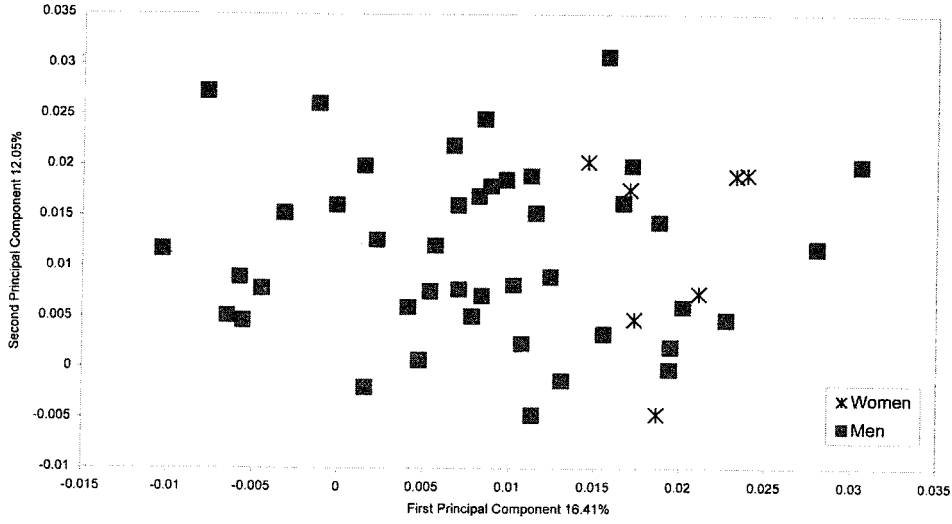
> Nay, that shall not serve your turn; that shall it not, in truth, la. Nay, I care not for such words. No, no. – And, my lord, he desires you that, if the King call for him at supper, you will make his excuse.
>
> (*Troilus and Cressida*, 3.1.72–5)

If he mouths the word 'truth' we can be sure the word 'not' is close by. He works upon others by qualifying and prevaricating: 'Faith, to say truth, brown and not brown', he says to Cressida of Troilus's colouring (1.2.92), trying, unnecessarily as it proves, to talk this blemish on Troilus's beauty out of existence.

Ulysses in the same play is at the other end of this spectrum. He has half as many *not*s as Pandarus, and a quarter the *I*s. At the extreme opposite to Pandarus is Warwick, notable for his use of *and*, with more than four instances in a hundred in his total dialogue. Warwick uses the conjunction to join events with a minatory inevitability: 'to London we will march, / And once again bestride our foaming steeds, / And once again cry "Charge!" upon our foes' (*3 Henry VI*, 2.1.182–4).

Looking at the disposition of the characters by way of the genre symbols in illustration 23 gives further insight into the differentiae that the method has identified as the strongest in this analysis of the commonest words in the largest Shakespeare characters. Characters from history plays (with black circles as markers) tend to be towards the left, characters from comedies (marked by shaded diamonds) towards the right. Nobles and kings cluster at the choric end, go-betweens and wisecrackers at the interlocutory one, and going along with these more extreme tendencies are the more mainstream characters of history plays and comedies. The Hal of *The History of Henry the Fourth* (labelled 'Prince Henry' in illustration 23[16]) is to the left, and Petruccio from *The Taming of a Shrew* to the right. Going against the trend is one of the three Falstaff characters, the Falstaff of *The History of Henry the Fourth* – the black disk to the lower right – a character from a history play in territory occupied mainly by characters from comedy. (The fact that

[16] His part in *2 Henry IV* is not long enough to qualify.



25 Characters identified by gender.

the three Falstaff parts constitute three of the fifty largest Shakespeare characters is itself of interest, a reminder that Shakespeare spent quite a large fraction of his total output in writing through this one vehicle.)

Ulysses from *Troilus and Cressida* (the shaded diamond to the top left) is the comedy character farthest to what we can call the history-play end. It is easy to see that this is because his is a choric role. He speaks for a collective rather than for himself, and about others rather than about himself. He makes well-developed pronouncements rather than involving himself in banter or altercation. There is not usually such a large role of this kind in Shakespearian comedy. Its presence is one of the many aspects that make this play unusual as a comedy.

Characters from tragedies (marked by hollow triangles) are in middling positions. Clearly the dialogue of tragedy overall is not sufficiently distinctive to make this one of the poles of Shakespeare's style in these terms. The graphs suggest that Shakespeare's generic range is better thought of as comedy versus history, rather than as comedy versus tragedy. This notion is supported by Hope and Whitmore's study of Shakespeare's writing as a 'Very Large Textual Object', based, like the present one, on computational methods, but with a much more mixed set of variables. Hope and Whitmore conclude that the contrast between history plays and comedies is the primary one to emerge from a linguistic study of the canon.[17]

To get a broad overview of the place of women characters as a group in this analysis, we can present the same results as in illustration 23, with labels this time reflecting gender instead of genre (illustration 25). The women characters are all in the right-hand half of the graph.

These characters do not often call on the grand coalitions implied by *we* and *our*; they speak more often of *I* and *me*; they respond to others' conversational gambits with *no* and *not*. There are seven women characters in all who speak more than three thousand words. The range in terms of the choric-reactive axis is from Portia (most choric) to Rosalind (most reactive), though this is in all not such a great range, only a little over a quarter of the span between Warwick and Pandarus.

[17] 'The Very Large Textual Object: A Prosthetic Reading of Shakespeare', paragraphs 21–33.

The vertical dimension of illustration 24 contrasts especially the masculine third person singular forms, and to a lesser extent the *you* forms, with *thou*. The users of *thou* and its partners are an interesting group: Romeo and Juliet, and then Prospero, King Lear, Petruccio and Titus Andronicus. There is no doubt that Shakespeare, along with the rest of his fellow-dramatists, and indeed his fellow speakers of English, used fewer of these forms as time went on. The replacement of *thou* forms with *you* ones is one of the most marked developments through the period of Early Modern English. The presence in the lower part of illustration 23 of Lear and Prospero, characters from the latter part of Shakespeare's career, goes against this trend. In his dialogue Prospero uses three instances of *thou* and *thee* to one of *you*. He addresses Miranda, Ariel and Caliban in this way, to an extravagant degree, one might say; this form is a handy shorthand for his miniature patriarchy, a tiny kingdom more or less willingly bound to its father-ruler. The circle inscribed by the pronoun is later extended to Ferdinand, and then to the rest of the Milanese court.

Lear's court is also a blend of family and kingdom, if more feudal in character. His use of *thou* is not quite as insistent as Prospero's, but it is still his dominant second-person form, with instances of *thou* and *thee* together outnumbering those of *you*. He bestows 'thou' on Cordelia, in his curses, and also in their reconciliation ('Thou art a soul in bliss' [4.6.39]). He calls Kent 'thou', as imagined recreant, and again as the masterless retainer Caius. The 'all-shaking thunder' (3.2.6) is addressed as 'thou', and so are most of the participants in the mock trial in scene 13 of the Quarto.

Illustration 23 shows Romeo exactly on a par on the vertical dimension with Juliet – their abundance of *thou*, *thee* and *thy* is a measure of the focus of their spoken parts on each other – but he is to the left of her on the horizontal dimension, slightly more authoritative and less reactive than she is.

At the top of the graph is Menenius from *Coriolanus*. He uses *he*, *his* and *him* frequently. Together they are more than three in a hundred of all the words he speaks. The main explanation for this is his focus on Coriolanus, whether as returning hero ('Is he not wounded? He was wont to come home wounded', 2.1.116–17), aspirant politician or treasonous exile. He turns only rarely to the *thou* forms, using *you* eight times as often as the corresponding *thou* and *thee*. Similarly, Claudius says 'he' often, 'thou' rarely; like Menenius, most of his instances of *he* refer to one person, in his case the absent and dangerous Hamlet ('he which hath your noble father slain', 4.7.4).

The displacement of the Falstaff of *The Second Part of Henry the Fourth* towards the northern border of illustration 23 and away from the southern one can be traced through his changed use of *thou* and *he*. To an extent this is a direct swap. In the first part Hal is most often present, and Falstaff's resilient hold over him is reinforced by addressing him as 'thou'. '[W]hen thou art king', he says, twice, within a few lines (1.2.16, 23). In the second part the Prince is more often absent, and becomes 'he'. This takes on a poignant quality in the final instance of the pronoun, in a speech to Shallow.

> That can hardly be, Master Shallow. Do not you grieve at this. I shall be sent for in private to him. Look you, he must seem thus to the world. Fear not your advancements. I will be the man that shall make you great.
>
> (*2 Henry IV* 5.4.75–9)

Overall Falstaff uses *thou* twice as often in the first part of *Henry IV* as the second, proportional to all the words he speaks, and *he* half as often. He is more an observer and commentator in the second part, losing his favoured position in the alternative royal court of Eastcheap, and being brought more into the daylight world of legal and military affairs.

Certainly it is an unfamiliar brand of characterization that we have been discussing here. One is led to talk not of imaginative or emotional life, or complex cognitive modes, but of vectors of interaction between characters, especially traced through pronouns, and the tricks of style with which they work on each other. This is a sociolinguistics of character. Roger Brown and Albert Gilman's classic study of *thou* and *you* forms is relevant here, with its commentary on instances in Shakespeare and their reflection of local questions of status and

relationship.[18] This work leads to a sense of the drama as a play of types, stock parts, defined this time not as villains, braggarts, wily servants and doting lovers, but more broadly by their place in a network of relationships. Among other things, Shakespearian characters speak for social purposes, to inform, persuade, control, seduce and sometimes to entertain. These purposes are reflected in their syntax and deixis and thus in frequencies of the very common words in their dialogue.

There are of course unending layers of complexity beyond this simple account. There are fifty characters in the analysis, but Shakespeare wrote more like a thousand parts in all. Characters change, and here they have been represented as a single static point. Shakespeare's original audiences heard the plays against a background of a repertoire of rival talents, and this is an important context for patterns of characters' dialogue, but here we have compared them only to other Shakespeare characters. The fifty words we have been considering represent a surprisingly large proportion of the total number of words these speak, around forty-five per cent, but only a tiny fraction of the total number of different words they use, more like two in a thousand. And of course, beyond the purely linguistic components are all the elements of action, setting, casting, acting and direction that make up a full picture of what it is to be Falstaff or Pandarus.

A study like this, then, starts with a drastic subtraction of all but a very few of the created and perceived materials that make for meaning in drama. It defines rigidly a small set of features to count and chooses one limited context in which to make comparisons among the results. The compensation is that this narrow set of variables is remarkably rich in information, since it offers access to patterns of syntax, of the structuring of language. Once the terms of the statistical analysis are defined, it misses nothing, and gives every item equal play. Its processes can be checked and replicated.

Most important, it presents us with some challenging propositions. Comedies and history plays are the generic poles of Shakespearian drama. It is the Falstaff of *The Second Part of Henry IV* who is the odd man out. Pandarus is defined by his *not*, Prospero by his *thou*, and Warwick by his *and*. In certain strictly defined terms, these observations are incontrovertible. They are discoveries, waiting to be made, recalling Falstaff's joking explanation for Worcester's taking to rebellion: it 'lay in his way, and he found it' (*1 Henry IV* 5.1.28). What bearing, if any, they have on anyone's understanding of Shakespeare is another matter. This is the province of readers or auditors, who are (mercifully) free to deploy in response to these propositions about the plays those many rich resources of interpretation of which the computer knows nothing.

[18] 'The Pronouns of Power and Solidarity', *Style in Language*, ed. Thomas A. Sebeok (New York, 1960), pp. 253–76.