



PROJECT MUSE®

The Hundredth Psalm to the Tune of "Green Sleeves": Digital Approaches to Shakespeare's Language of Genre

Jonathan Hope, Michael Witmore


Shakespeare Quarterly, Volume 61, Number 3, Fall 2010, pp. 357-390
(Article)

Published by The Johns Hopkins University Press

DOI: 10.1353/shq.2010.0002



➔ For additional information about this article
<http://muse.jhu.edu/journals/shq/summary/v061/61.3.hope.html>



The Hundredth Psalm to the Tune of “Green Sleeves”: Digital Approaches to Shakespeare’s Language of Genre

JONATHAN HOPE AND MICHAEL WITMORE

THE MERRY WIVES OF WINDSOR FEATURES SHAKESPEARE’S richest depiction of varieties of spoken English, but it also dramatizes—at a crucial point—the multiple processes of textual analysis, close reading, bibliographic description, and authorship attribution, all of which depend on the availability of written texts. Shakespeare’s textual analysts are Mistresses Ford and Page, each of whom has been the independent object of Sir John Falstaff’s romantic attentions. As Mistress Ford notes, Falstaff used his rhetorical skills to play the part of a gentleman: “he would not swear; praised women’s modesty; and gave such orderly and well-behaved reproof to all uncomeliness, that I would have sworn his disposition would have gone to the truth of his words” (2.1/620–24).¹

But when the knight moves from speech to writing, his by-the-book method of seduction can be compared by Ford and Page. The comparison provides a hilarious moment of parallel reading and recognition. Mistress Page quotes the lover:

“Ask me no reason why I love you; for though Love use Reason for his physician, he admits him not for his counsellor. You are not young, no more am I; go to then, there’s sympathy: you are merry, so am I; ha, ha! then there’s more sympathy: you love sack, and so do I; would you desire better sympathy? Let it suffice thee, Mistress Page,—at the least, if the love of soldier can suffice,—that I love thee. I will not say, pity me; ’tis not a soldier-like phrase: but I say love me. By me,

¹ All quotations from Shakespeare’s works are taken from Open Source Shakespeare, <http://www.opensourceshakespeare.org>, which attaches through-line numbers (TLNs) to the widely available electronic Moby Shakespeare, itself based on the one-volume 1864 Globe Edition edited by William George Clark and William Aldis Wright. We use the electronic text files of the Moby Shakespeare, with certain editorial preparations (removing speech prefixes, act and scene labels, and stage directions), for the iterative analyses described in this essay. References are cited by act and scene numbers, followed by a slash, then TLN. (Unless otherwise indicated, Web pages cited in this essay were accessed 1 August 2010.)

Thine own true knight,
 By day or night,
 Or any kind of light,
 With all his might
 For thee to fight, JOHN FALSTAFF"
 (2.1/572–86)

Falstaff's descent into doggerel shows him to be a better speaker than writer. His oral facility is well known, of course, but he misjudges his quarry, and this unintended iteration of the letter beyond its original addressee sets the comic reversal in motion. Mistress Ford knows the limits of her own charms ("What, have I scaped love-letters in the holiday-time of my beauty?") and so can readily call Sir John's tune: "I would have sworn his disposition would have gone to the truth of his words; but they do no more adhere and keep place together than the Hundredth Psalm to the tune of 'Green Sleeves'" (2.1/622–26). Writing both preserves what was written and allows that record to circulate beyond the intended recipient, which are the conditions for forensic analysis. The stage thus provides a model for a kind of iterative analysis made possible by a document:

Letter for letter, but that the name of Page and Ford differs! To thy great comfort in this mystery of ill opinions, here's the twin-brother of thy letter: but let thine inherit first; for, I protest, mine never shall. I warrant he hath a thousand of these letters, writ with blank space for different names—sure, more,—and these are of the second edition: he will print them, out of doubt; for he cares not what he puts into the press, when he would put us two.
 (2.1/632–41)

Mistresses Page and Ford do not need to see the comparison to recognize Falstaff's lewd intentions, of course: the formal textual analysis simply confirms what their separate close readings have already told them. Capture and comparison are enough. As readers and "users of texts," they provide us with a model for both the practice and benefits of the prosthetic analysis of language. As Renaissance thinkers knew very well, writing itself is a prosthetic: it allows us to overcome the physical limitations of the medium of speech and the psychological constraints of linguistic processing.² Similarly, digitally based research in the humanities expands the possibilities of iterative comparison glimpsed

² In Renaissance thought, writing is always an artificial technology—desirable and useful because it fixes man's transient words—but as the commonly made distinction between "words" (spoken) and "letters" (written) suggests, not conceived of as part of language itself. The Aristotelian formulation, repeated by almost all at the time, held that language (words) represented ideas (mental images). Writing, if it was mentioned at all, featured as a mere *representation* of words; see Jonathan Hope, "Ideas about Language in the Renaissance," in *Shakespeare and Language: Reason, Eloquence, and Artifice in the Renaissance* (London: Arden Shakespeare, 2010).

here, not just because more items can be stored and compared but because it is more productively indifferent to linear reading and the powerful directionality of human attention.

We begin with the *Merry Wives* scene of reading because it illuminates our larger purpose in this essay, which is to ask what it would mean to harness the potential of textual comparison this scene models and extend it to where human readers simply cannot go. We are interested in a kind of “iterative criticism” that links the wayward properties of documents—those provisionally bounded objects whose material form allows them to travel—with the inhuman power of arbitrary repetition proper to computation.³ What if everything Falstaff had ever said were transcribed—in a play, for example—so that it could be compared to every other utterance in the Shakespearean dramatic universe? And what if a congregation of canny readers, a klatsch of Mistress Pages or chorus of Shakespeare scholars, were ready with a list of comparisons they wanted to see made? Both the archive and the critical chorus are available to us now, either through the medium of print (the annals of scholarship) or through digitization, which provides us with primary texts whose contents can be manipulated and compared in ways that the original writer never intended.⁴ We might not be respecting the “original context of utterance” when we approach the Shakespearean archive in this way, but respect in criticism can mean a lot of different things. Critics frequently stray from the text in order to return.

Iterative criticism is a good name for our digital work with Shakespeare and the computer program DocuScope, because it makes explicit three assumptions about texts and our interactions with them: (1) texts must be “alienable” from

³ We recognize Stephen Ramsay’s “algorithmic criticism” in the genealogy of our own thinking on these matters; see his “Toward an Algorithmic Criticism,” *Literary and Linguistic Computing* 18 (2003): 167–74, and “Algorithmic Criticism,” in *The Blackwell Companion to Digital Literary Studies*, ed. Susan Schreibman and Ray Siemens (Oxford: Blackwell, 2007), 477–91, <http://www.digitalhumanities.org/companionDLS/> (accessed 2 March 2010). We like the word *iterative* because it links the nature of comparisons (which are arbitrary and repeated) to conditions of textuality, whose material supports always imply the possibility of circulation.

⁴ In the open commenting period, Lauren Shohet raised an important but complex question about the extent to which our work explains intersections between genres, modes, discourses. In our current work on the Very Large Diagram, with which we end this essay, we begin to confront the question of how we account for the patterning of variation we find in texts. How can we distinguish between patterning produced by genre, date, or author—or indeed character, acting company, or intended audience? This is a significant challenge and, as Hugh Craig notes in a recent essay, literary critics have often argued that the overlay of competing effects must make isolation of (for example) author effects impossible; see his “Style, Statistics, and New Models of Authorship,” *Early Modern Literary Studies* 15.1 (2009–10), <http://purl.oclc.org/emls/15-1/craistyl.htm>. It seems clear to us that, while explaining variation in texts is a complex activity requiring interpretation, as well as calculation, it is not an impossible one—theoretically or practically.

their original contexts in order to be compared, as in the case of Falstaff's letters; (2) the digital form these texts take is just a special case of a broader "juxtaposability" latent in language (Falstaff's words could also be overheard; that's what makes them words); and (3) comparisons are not self-instantiating: a critic or group of critics must always introduce a salient distinction for *any* repetitive technique to produce results (there must always be a critic or critics in iterative criticism; Mistress Page is not herself an algorithm, even if Falstaff's flattery is).⁵

Such work builds on the prosthetic notion of texts we began with, taking digital tools that count or aggregate features among texts to be *extensions* or *formalizations* of this prior technical augmentation of the human condition that is already found in writing.⁶ Our prosthetic is a computer program called DocuScope, created at Carnegie Mellon University in the late 1990s by David Kaufer and Suguru Ishizaki.⁷ We use this tool to provide detailed linguistic redescriptions of critically attested genres of Shakespeare's writing, particularly those of Heminges and Condell (Shakespeare's friends and editors) and the nineteenth-century Shakespeareans who argued for the existence of a distinct genre of Shakespeare plays, the so-called late plays or romances.⁸ In part, our published work has been designed to demonstrate that a phenomenologically based architecture for tagging English words—essentially, a collection of word buckets or dictionaries like DocuScope—could make *genre visible on the level*

⁵ The classic statement of this view of iterability as the *sine qua non* of textuality is Jacques Derrida's "Signature Event Context," which can be found in *Limited Inc.*, trans. Samuel Weber (Evanston, IL: Northwestern UP, 1998), 1–23.

⁶ The text itself might also be understood as a prosthesis or extension of or supplement to an underlying human limitation, a point made by David Wills in *Prosthesis* (Stanford: Stanford UP, 1995), 135. (We are grateful to Mark Burnett for pointing out this reference.) To the extent that genre extends the range of this prosthesis—by synchronizing expectations of readers and writers, for example—it too has prosthetic qualities. This "horizon of expectation," as Jauss understood, is by definition social, but it is also constrained by arrangements of objects and actors in the theatrical situation. Frances Teague's remarkable work on stage properties, which catalogues the object landscape of various genres, makes this level of constraint explicit. (We are grateful to Julian Yates for reminding us of the importance of this text.) See Hans Robert Jauss, *Toward an Aesthetic of Reception*, trans. Timothy Bahti (Minneapolis, U of Minnesota Pr, 1982); and Frances N. Teague, *Shakespeare's Speaking Properties* (Cranbury, NJ: Associated UP, 1991).

⁷ For DocuScope, see David Kaufer, Suguru Ishizaki, Brian Butler, and Jeff Collins, *The Power of Words: Unveiling the Speaker and Writer's Hidden Craft* (Mahwah, NJ: Lawrence Erlbaum Associates, 2004), x, xv. A discussion of how the program came to be designed and an early *précis* of its categories can be found at "Description of DocuScope," http://betterwriting.net/projects/fed01/dsc_fed01.html (accessed 3 March 2010).

⁸ See Jonathan Hope and Michael Witmore, "The Very Large Textual Object: A Prosthetic Reading of Shakespeare," *Early Modern Literary Studies* 9.3 (January, 2004): 6.1–36, <http://purl.oclc.org/emls/09-3/hopewhit.htm>; and Michael Witmore and Jonathan Hope, "Shakespeare by the Numbers: On the Linguistic Texture of the Late Plays," in *Early Modern Tragicomedy*, ed. Subha Mukherji and Raphael Lyne (Cambridge: D. S. Brewer, 2007), 133–53.

of the sentence. Thus, the intensive definitions we use to discriminate plays into groups—"comedies end in marriage," for example, or "the mood of these plays is similar"—can be tracked through a set of linguistic operations that take place in parallel to these perceptions but cannot themselves be consciously attended to. Nor, we would add, can one be reduced to the other.

Human consciousness is vectored, our attention scarce: we aggregate our perceptions into powerful impressions that interlace a vast number of comparisons, like the lightning-fast recognition of a family resemblance. Unlike the operations of consciousness, the operations of language—at least in drama—are more steady and deliberate. Shakespeare's language, we discovered, "does certain things" and does them repeatedly when a certain kind of story is being told. These linguistic doings are multiple, coordinated, and susceptible to statistical analysis. The critical prostheses we use to apprehend this other level of activity, then, are just extensions of the initial technologies of writing and comparison on display in our opening discussion of *Merry Wives*. Those prostheses include not simply the computer program itself, but also the linguistic, rhetorical, and cultural assumptions built into that program; the body of digitized texts we study; the codices and terminals that allow us to retrieve critical opinions from past and present writers; and the utilities of Skype and e-mail that we used to compose this paper between Kyoto, Japan, and Madison, Wisconsin. The result is a new kind of attention to texts and what they do with words, one that points us toward abstract representations of those linguistic activities only to return us to the texts themselves with renewed interest and questions.

GLOOP AND THE BANALITY OF DIGITAL READING: COMEDY AND HISTORY

We begin with an analogy based on a popular item of English cuisine: the pudding. Many English puddings feature a gloopy matrix in which something more substantial is intermixed, for example, a piece of fruit such as a plum. In our case, gloop is a useful substance to think with because it is analogous to the *linguistic* gloop that binds together the more spectacular items—the fabulous turns of phrase or memorable passages—that literary critics are likely to seek out and savor. As readers, we tend to ignore the ubiquitous gloop and prospect for the fruit, which means that we remain unaware of a large part of our reading experience. But if that matrix or gloop can be characterized by a machine, humans can return to the plums with a better sense of why they taste so sweet. Just as Page and Ford move from the forensic comparison of the identical letters to plotting their revenge on Falstaff, so digitally based research can provide a jumping-off point and even occasional guidance for human-based traditional reading.

Figure 1 is a plot of 776 pieces of Shakespeare plays. Each one contains one thousand consecutive words from a play (we discuss the reasons for chopping plays up so arbitrarily later). Each piece of text was subjected to rhetorical analysis by DocuScope, whose operations we will discuss in more detail below. The results of this analysis, which comprise frequency counts of just under one hundred linguistic categories, have been put through a complex but very common statistical procedure known as principal component analysis (PCA). The procedure makes comparisons between a large number of features within a population, allowing us to identify patterns of similarity and difference within the population based on correlating the presence and absence of features. If feature A is found in a group of the population, PCA asks if feature B is also present or predictably absent. PCA thus attempts to relate differences and absences within a population by making associations between them.⁹ These associations are expressed by placing members of the population at value points along a scaled principal component (PC). This procedure is good at making sense of complex relationships within large, complex populations—and, as a very excited statistician told us over lunch one day, Shakespeare’s language is one of the most complex and interesting populations around.

In this instance, the statistical package makes multiple comparisons between the relative frequencies of ninety-nine linguistic categories in the 767 thousand-word chunks of Shakespeare. It then combines the linguistic categories into PCs of highly positively and negatively correlated features, seeking to construct components which account for as much of the variation within the population as is mathematically possible.¹⁰ Each component is an answer to the questions “Are these bits of plays similar to each other?” and “Do the bits of plays form any groups with members of the group all sharing, or lacking, the same features?”

⁹ There are other techniques that could have been used to explore the variation in the data—one that has been employed recently by text analysts is Latent Dirichlet Allocation. We chose PCA because it is frequently used in statistics, which means its properties are well-known, and because it provides groupings of the plays that are often perfectly recognizable in terms of existing literary critical categories and discriminations. Other statistical techniques might produce “better grouping” but not be as easily tracked to ground-level language effects and strategies.

¹⁰ That we work on 1,000-word chunks of plays rather than whole plays is likely to strike readers as strange and arbitrary. Working with chunks of plays means that we identify features which are consistently used across the *whole text*: features used at a very high rate at just one point of a play will affect the score for just one or two chunks and will appear as outliers in a statistical analysis (of course, for some types of literary reading, we might be interested precisely in features which occur at a high rate at one point of a play). Furthermore, the mathematics of the statistics demands that populations be made up of more items than are being counted for. In this case, we are counting for ninety-nine linguistic categories in the thirty-six plays of the First Folio. “Chunking,” as this procedure is known, is a recognized and acceptable way to deal with this problem.

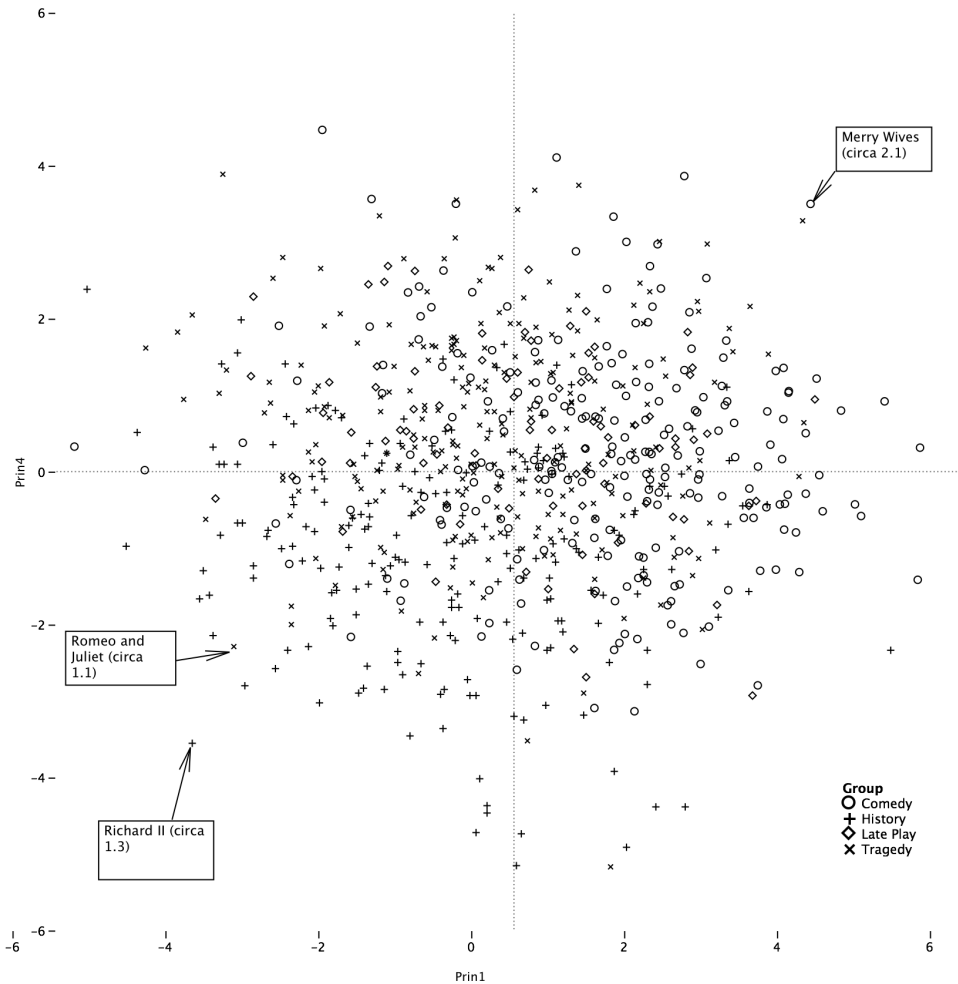


Figure 1: A total of 776 pieces of Shakespeare's plays from the First Folio, each piece consisting of 1,000 words, rated on two scaled PCs (1 and 4). The cumulative proportion of variation accounted for by the first four principal components is 12.33 percent, with component 1 accounting for 3.83 percent and component 4 accounting for 2.35 percent.

Figure 1 shows the results of running PCA on the DocuScope results from the fragments of the plays of the First Folio.¹¹ Our previous work established that there is a very clear linguistic distinction between Shakespeare's Comedies and the Histories,¹² and this figure confirms that finding on another level.¹³ In

¹¹ Color versions of figures can be found at <http://winedarksea.org/?p=816> and <http://digital.library.wisc.edu/1793/46265>.

¹² See Hope and Witmore, "The Very Large Textual Object."

¹³ In this paper, we capitalize the first letters of Comedy, History, Tragedy, and Late Plays when we refer to the linguistic features specific to these genres. Our aim in doing this is to emphasize the interpretive nature of these designations, made by the First Folio editors, and the

the figure, we have plotted the two PCs which account for most of the linguistic differences between Comedy and History: PC 1 (Prin1) on the horizontal axis, and Prin4 on the vertical axis.

We begin by noting that chunks of Comedy (marked by open circles) tend to score highly on *both* Prin1 and Prin4: scoring highly on Prin1 pushes them to the right half of the graph, while scoring highly on Prin4 pushes them to the upper half of the graph, with the result that most of the chunks of Comedy group together in the upper-right quadrant. Those readers with a traditional (or postmodern!) literary training may be tempted to focus on the outliers here—for example, one open circle at the extreme left of the graph. As we will see, these can be interesting, but for the moment remember that digitally based research is better at the gloop than the plums: the boring conformity, rather than the spectacular maverick. Conversely, chunks of History (marked by crosses) tend to be low on both PCs, resulting in a grouping of these at the lower-left quadrant of the graph. We could draw a diagonal line across this graph from upper left to lower right and leave most of the Comedy chunks above it and most of the History chunks below it. The statistical analysis tells us that there are highly significant, and consistent, linguistic differences between Shakespearean Comedy and History; but we should remember that all the analytic tools can “see” are 767 individual texts. The ascription of those chunks to the genres “Comedy” and “History” was done by the editors of the First Folio. Our analytic tools (DocuScope and PCA) identify linguistic similarities and differences in the population of text chunks. We have represented these visually and overlaid the Folio genre divisions. The extent to which the most significant linguistic similarities and differences in the population correlate with Renaissance genre divisions is striking.

So, one early claim of our work is that Shakespeare’s Comedies and Histories are linguistically distinct from each other. This distinctiveness can be shown statistically, and it is consistent. Let us try to unpick this claim as a way of demonstrating our methods, offering a critical understanding of iterative techniques and revealing the linguistic “gloop” or matrix of Shakespearean Comedy.

subsequent editors who called out the so-called “late plays” as their own category (*The Tempest*, *Cymbeline*, *The Winter’s Tale*, and *Henry VIII*; we follow the Folio editors’ generic designations of all the other plays). The language of Comedy, when it is referred to in this essay, is thus not the language of all comic writing *tout court*, but “comedy” as stipulated by the Folio editors. The printers of the First Folio influenced its generic scheme, adding printers to the list of historical actors responsible for the generic divisions we use in our analysis, a point made to us by Alan Galey. See also Jonathan Hope, “The Language of Genre,” in *Shakespeare and Language*, for an extended consideration of the function and context of the First Folio “catalogue” page, which makes these generic divisions visible. On the creation of the First Folio, see Charlton Hinman, *The Printing and Proof-Reading of the First Folio of Shakespeare* (Oxford: Clarendon Press, 1963); and Peter W. M. Blayney, *The First Folio of Shakespeare* (Washington, DC: Folger Shakespeare Library Publications, 1991).

First of all, what are Prin1 and Prin4? What does DocuScope count, the presence or absence of which is expressed by these scales? DocuScope is essentially a smart dictionary: it “reads” strings of characters looking for words and collections of words that it “recognizes.” When it encounters a word or phrase it knows, this string is counted. “Recognizes” means matches: DocuScope consists of a list, or dictionary, of over 200 million possible strings of English, each assigned to one of 101 functional linguistic categories called language action types (LATs). When DocuScope encounters a string it recognizes, the associated LAT is credited with one appearance. For example, “I” and “me” are strings which DocuScope assigns to the LAT “FirstPerson.” The occurrence of any one of them in a text is recorded as an appearance of the LAT “FirstPerson” (with one important caveat we will explain below).

Clearly, we are dealing with human interpretations and definitions based on a particular theory of how language works.¹⁴ DocuScope works in a mechanical manner in that it counts every string and every text it encounters in the same way, but the decision about what to count (what constitutes a functional string) and how to classify it (which LAT or higher category to put the string in) is not mechanical; ultimately, this is based on decisions made by the architect of DocuScope, David Kaufer, and these decisions are open to challenge.¹⁵ Digitally based research does not offer us the impossible dream of objective humanities research, but it does offer us the possibility of applying subjective humanities-based insights in a consistent way to test their applicability and utility across a large number of instances. Iterative criticism offers a way of being *consistently* subjective at a certain level of the analysis.

In DocuScope, a word can be counted only in one string; DocuScope always seeks to include a word in the longest possible string. So all instances of “I” are not automatically included in the LAT “FirstPerson”; those which occur with a tensed verb will be counted as “SelfDisclosure” because these strings are longer.

¹⁴ In this case, Halliday’s functional grammar; see M. A. K. Halliday, *Introduction to Functional Grammar*, 2d ed. (London: A. Hodder, 1994).

¹⁵ Early in our work we considered revising DocuScope’s string definitions and higher-level structure, since the program was developed for use on present-day English. This remains an option for the future, but we decided against it, largely for practical reasons. The initial construction of DocuScope took Kaufer almost a decade, with almost as much prior thinking and research; he might be justly referred to as the “Samuel Johnson of strings.” Even if we had a large amount of time to devote to a revision, it is by no means clear that much would be gained; because Kaufer did much of his string definition using the *Oxford English Dictionary* as a template, DocuScope deals with Early Modern English reasonably well (forms such as “thou” are included, for example). This is an example of a difference between traditional literary research, which tends rightly to be highly punctilious about choice of text, and digitally based research, where the volumes of data involved tend to make new preparation processes time intensive but also mean that low-level “errors” do not markedly affect the final results.

This raises an interesting issue: DocuScope was designed to allow the investigation of rhetorical effects on the assumption that different types of string have different types of experiential effects on readers. Implicit in the way it defines functional strings (a word joins the longest possible string, and only that string) is that individual words have one and only one functional effect on readers. In fact, we know from psycholinguistic research that linguistic effects can be multiple: words and sounds can “prime” for other words, for example.¹⁶ So DocuScope’s definition of string (the longest possible string, and only that string) may be necessary from a practical point of view, but on the level of individual words or clusters of words, its heuristic classifications are an oversimplification. This is the type of caveat we need to make explicit in digitally based research. Such a limitation does not render DocuScope’s findings meaningless. The patterns we have found so far are consistent across our work with DocuScope and make sense in terms of noniterative work on genre. But no investigative technique is without limitations: counting things is never simple.

Figure 2 is a graphical representation of the linguistic makeup of Prin1 and Prin4. We can think of it as repeating Figure 1, this time with the linguistic categories used in counting mapped onto the space rather than the chunks of plays. Once again, Prin1 is shown along the horizontal axis and Prin4 on the vertical axis. The data space is centered on point 0,0 at the graph’s origin, which represents a value of zero on both scaled PCs.¹⁷ From this point extend arrows, each one representing a LAT. The length of each arrow indicates the degree of loading that LAT has from being neutral for both graphed PCs. For example, a feature which appeared at the mean value for the whole sample would be graphed at 0,0, indicating that it played no role in distinguishing this group from any other. A feature such as “SelfDisclosure” has a long arrow to the right because it has a high positive loading on Prin1; play chunks high on Prin1 will have large amounts of “SelfDisclosure.” However, the arrow is horizontal because the LAT is neutral on Prin4—it plays no positive or negative role in ordering the plays as they appear along this scale.

As with Figure 1, we can imagine a diagonal line drawn from top left to bottom right, through the 0,0 point. Linguistic features characteristic of Comedy have long arrows above this line; linguistic features associated with History have long arrows below the line. With this in mind, we can start to pull out the

¹⁶ See, for example, Paula J. Schwanenflugel and Calvin R. White, “The Influence of Paragraph Information on the Processing of Upcoming Words,” *Reading Research Quarterly* 26 (1991): 160–77.

¹⁷ PCA was performed on the correlation matrix, which means that results are *scaled* and *centered*. This prevents the results from variables in which there is a lot of activity (for example, “description” strings) overwhelming those from rarer variables.

linguistic features which are statistically significant in making up the matrix of Shakespearean Comedy. A key point is that we are not only identifying presence, but also correlated absence. Shakespeare's Comedies are "high" on both Prin1 and Prin4 (this is why they cluster in the upper-right quadrant in Figure 1). They are characterized by those features which show positive scores on one or both of the axes.¹⁸ For example, "DenyDisclaim," "SelfDisclosure," "DirectAddress," and "FirstPerson" are all frequent in Comedy (and we will define and illustrate them in a moment). Conversely, Shakespeare's Comedies are characterized by a *lack* of those features which show strong negative scores on one or both of the axes—here, "Motions," "SenseProperty," "SenseObject," "Inclusive," and "CommonAuthorities."

Iterative research tells us that Shakespeare makes use of precisely those features he avoids in Comedy to constitute the matrix of History: the two variables "SelfDisclosure" and "SenseObject" are almost directly opposed. A loadings biplot like that shown in Figure 2 tells us that the use of one type of word (or string of words) seems to preclude the use of its opposite. This would be true of all the longer vector arrows in the diagram that extend from opposite sides of the origin.

For example, "LanguageReference," "DenyDisclaim," and "Uncertainty" strings are used in opposition to those classed under the LAT "CommonAuthority." If an item scores high on Prin4 (which most comedies do), it will be high in "Language Reference," "Uncertainty," and "DenyDisclaim" strings, while simultaneously lacking "CommonAuthority" strings. We can learn a lot by looking at this diagram, since once we decide that these components track a viable historical or critical distinction among texts it shows us certain types of language co-occurring in the process of making this distinction (for example, "this text is, or is not, a Comedy"). "DirectAddress" and "FirstPerson" tend to go together here (lower right), as do "Motions," "SenseProperties," and "SenseObjects" (upper left).

¹⁸ We chose PCs 1 and 4 from a much larger array of components that explain the variation in the Shakespearean corpus. The Tukey test shows that PCs 1 and 4 separate Comedies from Histories at a highly statistically significant level. Not all components identified by PCA separate genre groups: the remaining components track something else that criticism may or may not be able to name. We have found components that track authors, groups of authors, and generations of writers, for example, as well as other components that we are currently unable to explain. This excess of statistically viable patterns with respect to available critical categories presents a challenge to anyone offering a traditional structuralist interpretation of our results: if even a limited feature set produces many more patterns or principal components than can be explained with existing critical categories, how can one claim that a single category such as genre—isolated now as one of many statistical objects—is foundational? To say, as we do, that genre is a multivariate linguistic phenomenon visible at the level of the sentence is not to claim that it is the preeminent form of patterning in human symbolic activity.

Put another way, what this graph illustrates is what Mistress Ford detects in Falstaff's "disposition" and "words": both find a discrepancy among texts that do not "adhere and keep place together" any more than it is possible to set the hundredth psalm to the tune of "Green Sleeves." PCA shows us those things which consistently avoid each other, and those things which always accompany each other in texts.

LATs IN DETAIL: THE BUILDING BLOCKS OF SHAKESPEAREAN COMIC LANGUAGE

Iterative tagging techniques, then, can give us a statistical description of the language of Shakespearean comedy, pointing us toward features that characterize it in either their presence or absence. But few literary scholars will be comfortable simply accepting points on a graph as a *reading* of Shakespeare. Nor should they be. Digitally based research is not an end point: its findings need to be tested against the texts for two reasons. First, we must return to the text in order to ensure that meaningful items are being counted. In early work, we realized that speech prefixes were being counted by the program, producing artificially high totals in some DocuScope LATs. Second, we return to the texts because we are ultimately interested in how they are read by, and affect, humans. Digital approaches enable us to account for the effects of texts using new types of evidence. They do not create new textual effects, but rather allow us to describe their sources in a different way.

We have now supplied the names for what DocuScope "sees" when it looks for Shakespearean Comic language. But can we make sense of these results in terms of our own reading of how Shakespearean Comedy works? Many LATs whose presence is characteristic of Comedy code for high levels of verbal interactivity: "RefuteThat" and "DenyDisclaim" both pick up active negations within texts. "RefuteThat" strings tend to make assertions using a negative and typically consist of subject plus copula verb plus negative judgment ("it's nonsense") or pronoun plus refutative verb ("I deny that"). They imply or assert explicitly that another statement is false ("but the reality is"). With Shakespeare texts, DocuScope most frequently counts in this category items such as "but," "yet," "not so," "rather," "revenge," "it is not," "I will never," "will not let," and "for all that." "DenyDisclaim" strings strongly imply a previous statement that is being negated ("there is no conspiracy"). DocuScope most frequently counts in this category items such as "not," "no," "nor," "never," "there is not," "it cannot be," "it is not so," "this is not," and "cannot chose but."

Similarly, "DirectAddress" categorizes strings which challenge or directly call for attention from an addressee. A frequent form is the second-person pronoun ("you," "thou") or the same pronoun plus modal plus verb ("you should con-

sider") and constructions such as "let us," "you shall," "I must," and "how say you." This gives us an initial picture of a Shakespearean Comic language that exists mainly between individuals jointly involved in the production of discourse, actively exchanging opinion and information about the world, and actively disputing other versions of the world.

We can extend this picture by looking at the following extract from *Twelfth Night* (Figure 3), showing where DocuScope registers the LATs plotted in Figure 2. Remember that these are some of the linguistic elements which allow us to differentiate Shakespearean Comedy from the other genres purely on the basis of statistics. To the LATs detailed in the previous paragraphs (here, "Deny-Disclaim" and "DirectAddress"), we can add "FirstPerson" ("I," "me," "my"); "SelfDisclosure" (typically involving a first-person pronoun with a verb implying some form of revelation: "I think," "I am," "my passion"); and "Uncertainty" (typically expressing the subjective nature of declarative statements: "seem," "perhaps," "things"). Note that these features can be associated with rapid-fire interaction: they cluster at the start of this extract as Olivia and Viola exchange single-line speeches and they drop off in frequency as soon as Olivia begins a prolonged speech ("O, what a deal of scorn"). We need to bear in mind when reading this extract that it is presented as *typical* of Comic language. This is what Shakespeare consistently and persistently does in the Comedies: it is "special," in that it is worth quoting, only in as much as it is normal for the Comedies. Can we link these statistically established linguistic patterns with our critical sense of how Comedies work?

Twelfth Night has three interesting plot devices—a set of identical twins, a shipwreck, and a disguise, all of which introduce a high degree of unintentional confusion into the action, driving it forward. In a plot driven by accident and what you might call "congruent misunderstanding" (when two people do not realize that they are speaking at cross-purposes), you expect to find a lot of back and forth between characters as they synch up their erroneous suppositions (which is funny in and of itself) and then more back and forth as they backtrack in order to rehearse why they did not understand what was going on when they were so deeply engaged with one another.¹⁹ The first thing we notice about this exchange is that it involves an extended miscommunication, culminating in the wonderful line "I am not that I play" (1.5/478).

The doubled first person is emblematic of the doubling of Viola's person in Cesario (or in Olivia's apprehension of Viola as Cesario), and we can see from the underlined strings in Figure 3 how the comic jousting over identity results in a high frequency of "FirstPerson" and "SelfDisclosure" strings. The

¹⁹ In one of Shakespeare's early comedies, this process is called "sympathized . . . error" (*Comedy of Errors*, 5.1/1842).

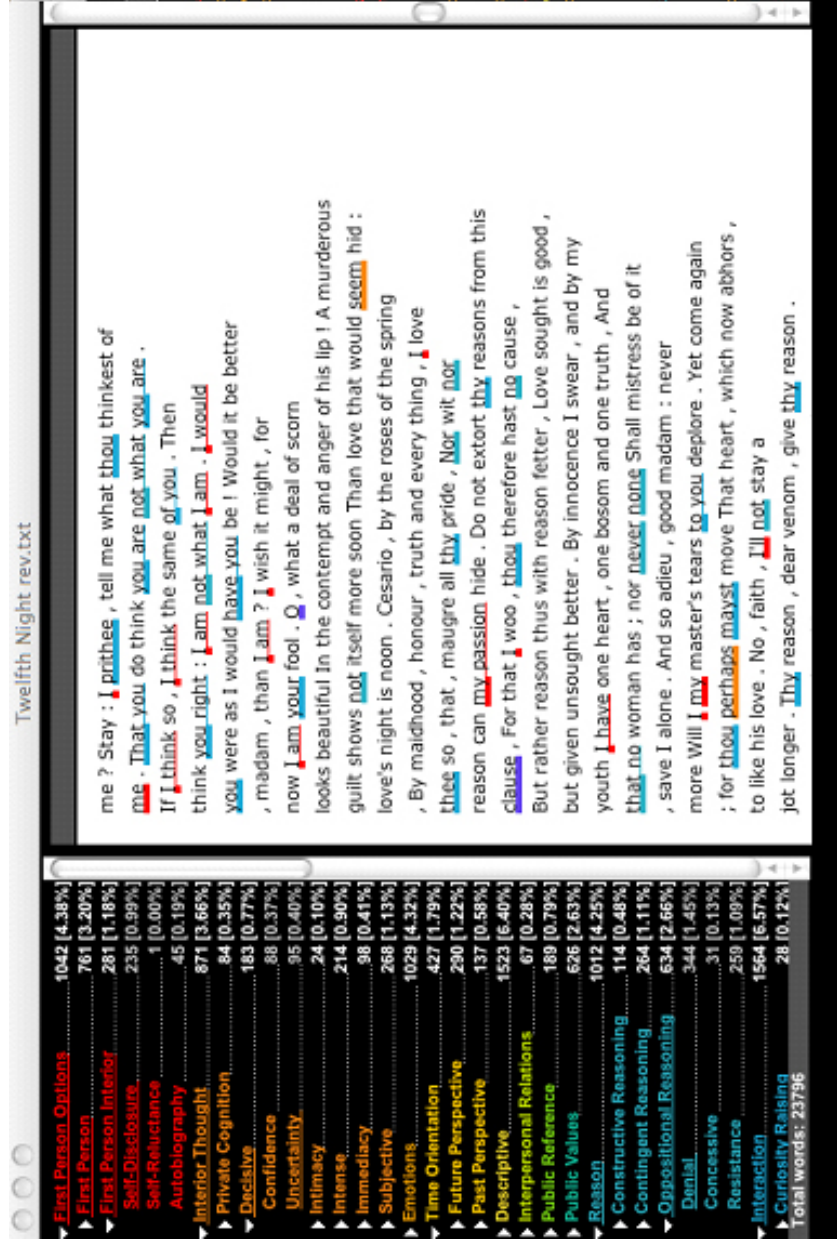


Figure 3: DocuScope screenshot of exemplary Comic strings from *Twelfth Night*, 3.1.

other type of strings that characterize Comedy are "Uncertainty," "Denial," and "DirectAddress." In context, these LATs would seem to support the idea that comedy is built on a linguistic matrix of dialogue that is (in a certain sense) talk to another person about talk. "Denials" are required to call into question previous statements, "DirectAddress" strings identify the origins of those statements, and "Uncertainty" strings mark the speaker's inability to see those statements through to a communally shared reality.

The quick trading of *I/you* and *my/your* strings in Comic dialogue suggests a world in which predicates are attached to subjects from two, and only two, points of view. This is not a universe of one; nor is it a crowd. It is not surprising that Comic plotting, built as it is on sexual pairings, would favor this type of bivalent, perspectival tagging of action by speakers. But there is something else going on here. Olivia is trying to make something happen in this exchange. She says, "do not extort thy reasons from this clause," and earlier, "I would you were as I would have you be!" (3.1/1392, 1381). The "thy" and "you" are important because the speaker is trying to create or assert a particular interpretation of how these two individuals relate to one another (and the words exchanged between them). The essential drama in this situation is the asymmetry of desire that obtains between the two characters, an asymmetry that keeps Viola from assenting to Olivia's advances. That resistance is actually what forces Olivia to make these statements that are rich with *I/you* and *me/my*, since she uses these words as anchors for a broader interpretation that does not yet obtain. She really wants to say *we*. And Cesario doesn't, so they remain in *I/you* dialogue.

To a certain extent, then, we can see what in the language makes it Comic in the eyes of DocuScope. But as we mentioned at the outset, definitions within PCA are built on absences, as well as presences: what's missing from Shakespeare's comedies, statistically speaking, are strings that make reference to the physical world. The entire component that characterizes Comedy, then, is one in which "FirstPerson," "SelfDisclosure," "DirectAddress," "Uncertainty," and "Denial" strings are mutually elevated from the mean score of all plays, while "Motions," "SenseProperty," "SenseObject," "Inclusion," and "CommonAuthority" strings are (simultaneously) below the mean.

The first three of these LATs are concerned with the description of the physical world. "Motions" track spatial relations via strings such as "lie," "fly," "draw," "walk," "fetch," "shake," "throw," "touch," "stir," "blows," "move," "close," "blow," "carry," and "rise." "SenseProperties" strings are typically made up of adjectives (frequently attributive) such as "sweet," "old," "young," "little," "long," "light," "sweet," "cold," "sound," "hot," and "heavy." "SenseObjects" are concrete nouns: "hand," "blood," "eyes," "heart," "the king," "bear," "tongue," "head," "eye,"

“sword,” “house,” “face,” “hands,” “bed,” and “gold.” The last two key us into the consensual, communal, public world of History, as distinct from the typical dyadic, oppositional one of Comedy. “Inclusion” is marked by “our,” “us,” “Our,” “of our,” “together,” “we have,” “to our,” “in our,” “ourselves,” and “that we.” “CommonAuthority” is shown by such items as “lord,” “God,” “Lord,” “unto,” “lords,” “duke,” “majesty,” “Duke,” “royal,” “highness,” “warrant,” “gods,” “command,” “he that,” and “sovereign.” (These strings are further illustrated in Figure 7 below, which shows a particularly typical History fragment from *Richard II*.)

We should pause at this point to note that there is no obvious linguistic reason why a text should not have high frequencies of “FirstPerson,” “SelfDisclosure,” “Uncertainty,” “DirectAddress,” and “Denial” strings while also having high frequencies of “Motions,” “SenseProperty,” “SenseObject,” and “Inclusion” strings. We can imagine a text or genre where characters argue energetically about the nature of the physical world around them, exchanging alternative and opposing theories about things—perhaps Tom Stoppard’s *Arcadia*. But this is not what happens in Shakespeare. In Shakespeare’s dramatic works, for some reason, rapid personal exchange and argument seem to preclude an interest in the physical world, and vice versa.

Perhaps a linguist could explain this pattern as a general feature of the language to show that our language can only “bend” in certain ways, making it quite difficult to use a lot of concrete descriptive nouns and words describing motion or changes in states of objects while simultaneously juggling lots of *I/you, my/your* strings. But this would not be enough of an explanation for us. We need to say why this type of language pattern—whether or not it is constrained by limits in our grammar, cognition, or underlying semantic maps—coincides with genre classifications made by discriminating humans (Heminges and Condell, Shakespeare’s editors). The overlap is what is most interesting, even if that overlap suggests some underlying constraints on language use and narrative that we have not really considered in literary critical work.

We can now offer a preliminary hypothesis. Shakespeare writes Comedies in which characters, sometimes quite perversely, find the wrong way to the ones they love. Often it is chance or an onstage helper who sorts this out. Shakespeare is actually quite reserved when it comes to showing love as naturally progressing through its obstacles unassisted. But given that in the initial stages of courtship Shakespearean lovers almost never meet and join in a perfectly symmetrical way—they don’t start out as stones set in an arch, leaning perfectly on a key-stone—we should expect this asymmetry to show itself in the language. Where does it show up? It appears when a resistant individual, a “you,” prevents another “I” from arriving at an interpretation of a relationship that might be referred to as a “we” before others. Let’s call this the “resistant-you” hypothesis. Linguisti-

cally, the effect manifests itself in the assertion of the self ("FirstPerson") and the rejection of suggested mental and emotional realities ("DenyDisclaim").

THREE PLUMS: *OTHELLO*, *RICHARD II*, AND *ROMEO AND JULIET*

Once we establish reliable descriptions of what Shakespeare does most of the time, we can look at some results which a statistician would probably class as outliers, but which a literary critic is likely to pick out as the most interesting. Again, it is worth stressing the absolute difference in approaches here. Statistics expects outliers in any population of results (a set of results with no outliers is likely to be viewed suspiciously). Crucially, statisticians are generally not interested in such results, and may even employ measures to exclude them from the analysis. This makes sense; if you are trying to establish what Shakespeare's Comic language is typically like, including a play nominally termed a Comedy but in which the writer behaves (for whatever reason) as if he were writing a History may skew your results and leave you with an unclear picture of comic language. Perhaps, too, you would be put off by a play that was in statistical terms an *extreme* example of Comedy, which manifested its signal features in ridiculous abundance. Outliers can be distractions in statistics. In literary studies, on the contrary, they can strike us as exceptions that illuminate a convention or shared expectation.

To a literary critic, then, a Shakespearean experiment, even a failed one, is highly interesting in itself—and worthy of particular study. We can now begin to see the need for interchange between digitally based and more traditional research techniques. There is no basis on which a purely iterative or algorithmic method can distinguish between genuinely interesting outliers (which are significant in a nonstatistical sense) and the expected but meaningless statistical blips any data set includes. Only traditional reading can identify those outliers with something to tell us about Shakespeare's language. But iterative techniques applied to a digitized text can call attention to outliers, and potentially tell us more than "what we already know" from our own reading.

One does not always need an outside prompt like statistics to begin exploring counterintuitive ideas about how literary or dramatic texts work. Among literary critics, some very distinguished readers (or auditors) of Shakespeare's plays have argued that he sometimes builds one type of play on the foundations of another. In the late 1970s, for example, Susan Snyder argued that a comic "matrix" underlies Shakespeare's tragedies. That is, Shakespeare built some of his tragedies—*Othello*, in particular—on structures that would ordinarily be employed in comedy and in doing so heightened the emotional effect of downturn in the plays when things deteriorate.²⁰

²⁰ See Susan Snyder, *Shakespeare: A Wayward Journey* (Newark: U of Delaware P, 2002), 29–45, and *The Comic Matrix of Shakespeare's Tragedies* (Princeton: Princeton UP, 1979), 70–74.

There is thus a certain, almost structural, irony to *Othello*. Some of what you see happening on stage seems to evoke the expectations of comedy (and its happy conclusions), but what eventually transpires is the opposite. While this may sound emotionally perverse, linguistically speaking it is exactly what Shakespeare was up to in this play, and it should not be surprising that a reader as careful and informed as Snyder figured this out. One of the most interesting consequences of this reading is that we begin to think of genre as something dynamic: a transaction between a spectator and a company that is full of false starts, head fakes, and allusive gestures. Perhaps rather than a recipe or essence, theatrical genre is really an oscillation between certain generic possibilities at a given moment in time. The insight that genre is comparative or differential is not, in and of itself, new: it is implicit in Fowler's analysis of genre in terms of Wittgensteinian family resemblance, an approach carried forward by Barbara A. Mowat in her analysis of Shakespearean romance.²¹ What is new is the idea that this dynamic difference is legible at the level of the sentence: that genre goes all the way down to where an author plants his or her feet in the ground, and can be tracked like a dance step if we keep our eyes on the floor rather than on the gestures of the hands and upper body. What DocuScope finds is something like the massive vertical integrity that holds among differing layers of language, from the most particulate (pronouns, pronoun-verb combinations) to the more semantic (words drawn from particular fields of use, such as motion or description), to the transactional units of plot (go here and do this) all the way to the level of imagery that critics are often drawn to as "emblematic" of some larger experience of the play's structure.²²

But however we choose to think about genre, it is safe to assume that we never encounter specimens that are "pure to type." As with the case of illustrators of botanical species, the artist may have one or many individual specimens at

²¹ See Barbara A. Mowat, *The Dramaturgy of Shakespeare's Romances* (Athens: U of Georgia P, 1976), 36, 69, and "'What's in a Name?' Tragicomedy, Romance, or Late Comedy," in *A Companion to Shakespeare's Works*, vol. 4, *The Poems, Problem Comedies, Late Plays*, ed. Richard Dutton and Jean E. Howard (Malden, MA: Blackwell, 2003), 129–49, esp. 134. Mowat credits the adaptation of Wittgensteinian "theory of family resemblance" to genre theory to Alastair Fowler's *Kinds of Literature: An Introduction to the Theory of Genres and Modes* (Cambridge: Cambridge UP, 1982).

²² This vertical integration has been confirmed by experiments performed by Matt Jockers at Stanford, who uses the most frequent words in the Shakespearean corpus—what linguists call "function words"—to produce genre groupings that are remarkably similar to the ones we have produced with DocuScope. See Matthew L. Jockers, "Machine-Classifying Novels and Plays by Genre," 13 February 2009, <http://www.stanford.edu/~mjockers/cgi-bin/drupal/node/27> (accessed 26 July 2010). See also the remarks on the use of function words in author attribution in Stanley Wells and Gary Taylor with John Jowett and William Montgomery, *William Shakespeare: A Textual Companion* (New York: W. W. Norton, 1997), 80–89.

hand, but the question is always whether or not to “idealize” or “mix” the specimens in order to depict the ideal type. Such types do not really occur in nature. Or if one settles on a particular example as the ideal then it will be, strictly speaking, a class of one, since all other specimens will deviate slightly from the illustrated example.

When we turn to the population that is mapped by DocuScope, we immediately see that *Othello* is not true to type. *Othello* is placed, as perhaps Snyder would have predicted, in the sector where many Comedies gather. We repeat Figure 1 with a slight difference; all of the plays are shown as small crosses—we are using the same PCs (Prin1 and Prin4)—but *Othello* is now highlighted as a series of empty circles (Figure 4).

So, is DocuScope “right” in calling *Othello* a Comedy? Was Snyder “right” in saying that the play was built on a comic “matrix”? Is there anything to be learned from the fact that DocuScope and a particularly distinguished critic agree on where *Othello* belongs? We should begin thinking about these questions by looking at specific passages. Below is an exchange between Othello and Iago, a dialogue between two individuals that looks a lot like the Comic exchange we examined from *Twelfth Night*. This is the beginning of what some critics have called the Othello’s seduction by Iago, a seduction that culminates in Othello’s kneeling before his former servant in a new misogynistic alliance.

- | | |
|---------|--|
| IAGO | I am glad of it; for now I shall have reason
To show the love and duty that I bear you
With franker spirit: therefore, as I am bound,
Receive it from me. I speak not yet of proof.
Look to your wife; observe her well with Cassio;
Wear your eye thus, not jealous nor secure:
I would not have your free and noble nature,
Out of self-bounty, be abused; look to't:
I know our country disposition well;
In Venice they do let heaven see the pranks
They dare not show their husbands; their best conscience
Is not to leave't undone, but keep't unknown. |
| OTHELLO | Dost thou say so? |
| IAGO | She did deceive her father, marrying you;
And when she seem'd to shake and fear your looks,
She loved them most. |
| OTHELLO | And so she did. |
| IAGO | Why, go to then;
She that, so young, could give out such a seeming,
To seal her father's eyes up close as oak—
He thought 'twas witchcraft—but I am much to blame;
I humbly do beseech you of your pardon
For too much loving you. |

OTHELLO I am bound to thee for ever.
 IAGO I see this hath a little dash'd your spirits.
 OTHELLO Not a jot, not a jot.
 IAGO I' faith, I fear it has.
 I hope you will consider what is spoke
 Comes from my love. But I do see you're moved:
 I am to pray you not to strain my speech
 To grosser issues nor to larger reach
 Than to suspicion.
 OTHELLO I will not.
 IAGO Should you do so, my lord,
 My speech should fall into such vile success
 As my thoughts aim not at. Cassio's my worthy friend—
 My lord, I see you're moved.

(3.3/1845–81)

This is yet another passage in which *I/you* interaction (characterized by “First-Person” and “Interaction” strings) occurs at the expense of concrete description (Figure 5). This is what, statistically speaking, pushes the passage into the zone normally occupied by Comedy. If there is a Comic matrix here—and not just in the happy setup of the early acts—it is the continued stance that allows a “withholding speaker” (Iago) and an eager listener (Othello) to push back and forth on one another. Othello here plays the role of Olivia in *Twelfth Night*, trying to delve further into the thoughts of his interlocutor (which keeps the *I/you* and *I/thee* pronouns coming) while Iago is a sort of Cesario, refusing to give the speaker something he wants (and in doing so, goading the speaker on). The parallel is perverse, but it shows that a very different emotional trajectory can take shape on a similar linguistic footing, much as a dancer can perform different upper-body movements on a similar footing or stance.

The next passage deepens the analogy in disturbing ways. In this scene from Act 4, close exchanges between Othello and Desdemona are structurally similar to those of the recognition scene in *Twelfth Night*. Notice how Othello's complaints echo the type of complaints one hears from a Petrarchan lover, although they emerge from an alienation and tragic emotional development that DocuScope cannot count in its perpetual “now” (Figure 6).

OTHELLO [TO EMILIA] Some of your function, mistress;
 Leave procreants alone and shut the door;
 Cough, or cry “hem,” if any body come:
 Your mystery, your mystery: nay, dispatch. [EXIT EMILIA]
 DESDEMONA Upon my knees, what doth your speech import?
 I understand a fury in your words.
 But not the words.

- OTHELLO Why, what art thou?
 DESDEMONA Your wife, my lord; your true
 And loyal wife.
 OTHELLO Come, swear it, damn thyself
 Lest, being like one of heaven, the devils themselves
 Should fear to seize thee: therefore be double damn'd:
 Swear thou art honest.
 DESDEMONA Heaven doth truly know it.
 OTHELLO Heaven truly knows that thou art false as hell.
 DESDEMONA To whom, my lord? with whom? how am I false?
 OTHELLO O Desdemona! away! away! away!
 DESDEMONA Alas the heavy day! Why do you weep?
 Am I the motive of these tears, my lord?
 If haply you my father do suspect
 An instrument of this your calling back,
 Lay not your blame on me: If you have lost him,
 Why, I have lost him too.
 OTHELLO Had it pleased heaven
 To try me with affliction; had they rain'd
 All kinds of sores and shames on my bare head,
 Steep'd me in poverty to the very lips,
 Given to captivity me and my utmost hopes,
 I should have found in some place of my soul
 A drop of patience: but, alas, to make me
 A fixed figure for the time of scorn
 To point his slow unmoving finger at!
- (4.2/2770–2803)

“[W]hat art thou?” Othello asks. And Desdemona answers, “Your wife, my lord; your true / And loyal wife.” Like Viola declaring who she is to Sebastian in *Twelfth Night*, Desdemona asserts who—not what—she is in the face of something like a disguise, forced upon her by Iago’s accusations. She is trying to puncture the veil of Othello’s illusion. Yet, instead of the gladness of recognition, we get a strange catalogue of personal suffering, a lover’s complaint over a loss never really suffered. This could, in other words, be a catalogue of suffering that has ended, but instead Shakespeare writes it as a kind of torment that has just begun. Linguistically, it contains all of the strings that DocuScope sees as key in clustering this play together with others we would call Comedies. But comic it is not.

What fascinates us about passages that are antigereneric in type is that they show the deep flexibility of anything we might call a structure or matrix on the linguistic, statistical level. There is no “essential structure” of comedy here, since Tragedies can exploit the same postures or stances that Comedies use to comic effect. This is something a machine can “see,” but a sensitive critic can see it as well. Yet a critic might not describe that matrix in the way that we have here—as

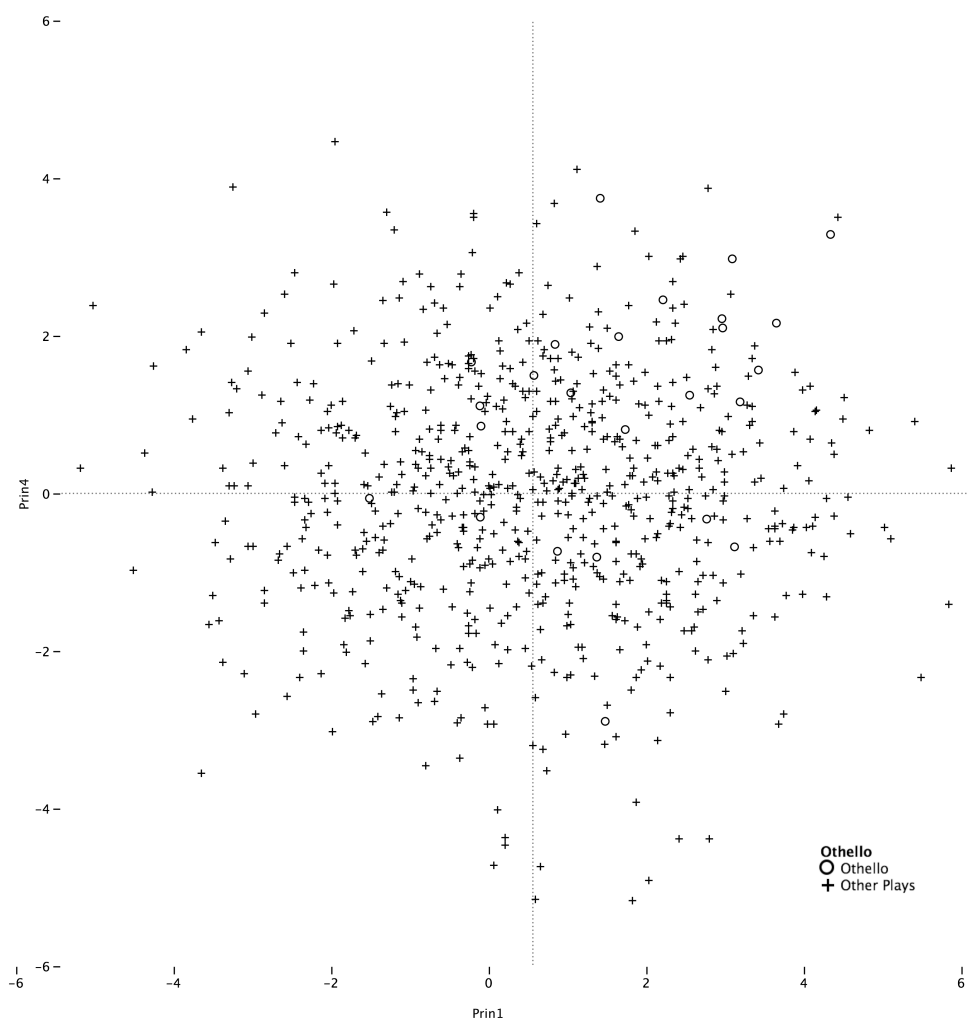


Figure 4: A total of 767 1,000-word pieces of the Folio plays rated on scaled PCs 1 and 4. This image is the same as Figure 1, except that all the plays are displayed as small crosses, with the exception of *Othello*, which is displayed as empty circles and collects mostly in the upper-right-hand quadrant where the Comedies tend to cluster.



Figure 5: DocuScope screenshot illustrating exemplary Comic strings from *Othello*, 3.3. In the online, full-color version of this figure, LATs were distinguished by underlining in different colors, a distinction lost here.

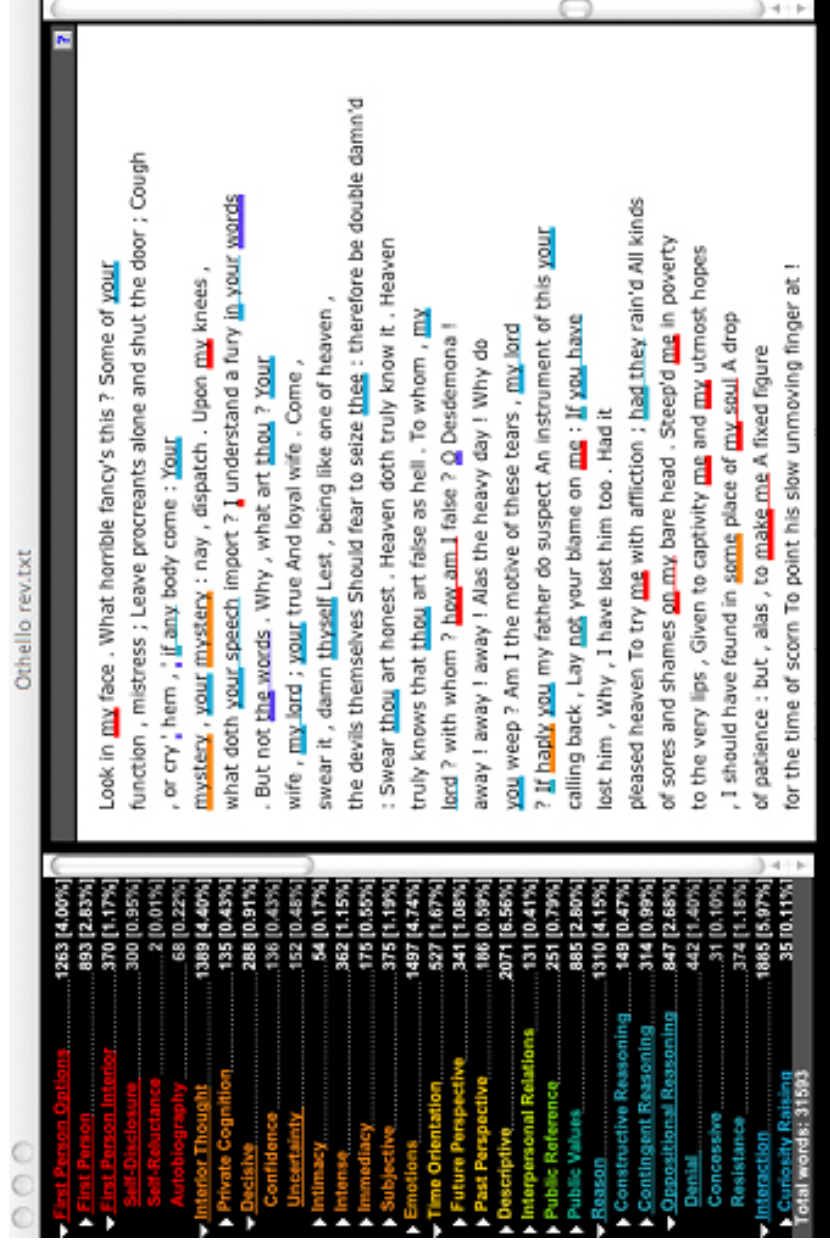


Figure 6: DocuScope screenshot illustrating exemplary Comic strings from *Othello*, 4.2.

a collection of present and absent linguistic tokens classed by type—and this is where DocuScope begins to throw up new questions about the play, genre, and reading. When Snyder said that *Othello* has deep affinities with Comedies, was she reacting to the linguistic cues described above? Are these features co-occurrent with the more intensive features that she did read for? What is the nature of this co-occurrence or shared footing of particular linguistic patterns and generic types? And how much antitypical language can there be in a play of a given type—for example, how much much “Comic” language can a Tragedy like *Othello* tolerate? Finally, what does this type of linguistic borrowing say about the ways in which genre is staged, cued, and self-consciously manipulated by authors? Would it be self-defeating to say that *Othello* is a good Tragedy because it uses Comic linguistic features to novel effect? This latter claim would, of course, be a matter of interpretation. But it is possible, as we saw in the chunking experiment above, to see how often parts of a particular play stray into other generic territories, and to quantify just how convergent certain parts are with a given antitype.

Consider our original graph of the two PCs that are most effective at separating Comedies from Histories when the plays are divided into thousand-word chunks (see Figure 1 and front cover). An extreme specimen of History writing appears in *Richard II*, circa 1.3, in the lower-left-hand quadrant. Strings responsible for pushing this piece of the play down and to the left (which accounts for its low rating on PCs 1 and 4) are highlighted (Figure 7). We see the formal settings of royal display, a herald offering Mowbray’s formal challenge, exactly the kind of stage transaction we would expect from a History play, in which the rituals of court and aristocracy are central to the dramatic action. Words underlined are categorized under the LATs “SenseObjects,” “SenseProperties,” “Motion,” “Inclusiveness,” “CommonAuthority.” Chairs, helmets, blood, earth, gentle sleep, drums, quiet: we don’t think of history as the genre of objects and adjectives, *but linguistically it is*. “Inclusive” strings are perhaps less surprising, given our previous analyses. We expect kings to speak about “our council” and what “we have done”; of course, “we” represents a presumed community that cannot be assumed in the back-and-forth dialogue of frustrated love in the comedies. Indeed, such an inclusive plurality is exactly what was missing from the Comedies, dominated as they are by the first person singular.

But look now at the tragedy chunk that scores lowest on Prin1 and Prin4—the item up and to the right of the *Richard II* fragment in Figure 1. This is the opening of *Romeo and Juliet*, the most historical piece of tragic writing that Shakespeare produced, according to this linguistic analysis. Here again, we give the marked-up Moby Text as we view it in DocuScope, calling attention to the strings pulling this piece of the text into the historical quadrant (Figure 8). This scene is the linguistic cognate of the one from *Richard II*: a voice of authority (the Prince) is called on to

adjudicate a conflict between nobles. We have the same preponderance of sensuous objects being given particular qualities—Tybalt is fiery, defiance is breathed, we hear of heads, windows, thrusts, and blows—and this reality of things and persons is counterpoised by that of the aristocratic community implied in the Prince's "we." Words like "citizens" and "civic" attest to the presence of communally sanctioned authorities rather than the private passions that govern love. Furthermore, we see a lack of words indicating uncertainty ("perhaps," "seems") and acts of self-disclosure ("I am," "I have") prevalent in the Comedy fragments examined above. But this linguistic convergence suggests a different kind of overlap as well. As the Oxford editors suggest, *Romeo and Juliet* and *Richard II* were written in close proximity to one another, probably in the year 1595.²³ It is possible, then, that as he was setting out to write his second Tragedy in a career filled with successful History plays, Shakespeare used the type of scene and language that was very familiar to him in his previous plays, but this time as the starting point for a works that would develop along different generic lines.

FILIATION, STRUCTURED ACCIDENTS, AND THE VERY LARGE DIAGRAM

This last instance of cross-generic filiation in Shakespearean writing introduces an interesting possibility that we can now begin to explore in iterative criticism: that Shakespeare deliberately writes across generic lines at different points in his career, and that individual plays arc into and out of zones of generic intelligibility while occasionally leaning out over the waters, so to speak, in order to incorporate material that is generically contrary to type. To what degree is any one of Shakespeare's genres tolerant of such atypical generic behavior, and how does this tolerance expand or contract with reference to the broader literary or textual field? If genre is as deeply embedded in a linguistic matrix as we believe it is, how might certain types of writing "travel" over time or even geographically, in something like a mobile, dynamic field of writing?

Posed at the level of corpus-wide linguistic features, such questions are not entirely foreign to literary studies and have been broached by Franco Moretti, Robin Valenza, and Brad Pasanek.²⁴ Perhaps the deeper significance of our find-

²³ Wells and Taylor, 117–18.

²⁴ Franco Moretti, *Maps, Graphs, Trees: Abstract Models for a Literary Theory* (London: Verso Books, 2005); Robin Valenza, "How Literature Becomes Knowledge: A Case Study," *ELH* 76 (2009): 215–45; and Brad Pasanek and D. Sculley, "Mining Millions of Metaphors," *Literary and Linguistic Computing* 23 (2008): 345–60. See also J. F. Burroughs, *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method* (Oxford: Clarendon Press, 1987); and Douglas Biber, Susan Conrad, and Randi Reppen, *Corpus Linguistics: Investigating Language Structure and Use* (Cambridge: Cambridge UP, 1998). One of us (Witmore) is engaged in a longitudinal study of Victorian novels using DocuScope with Sara Allison, Ryan Hauser, Matt Jockers, and Franco Moretti.

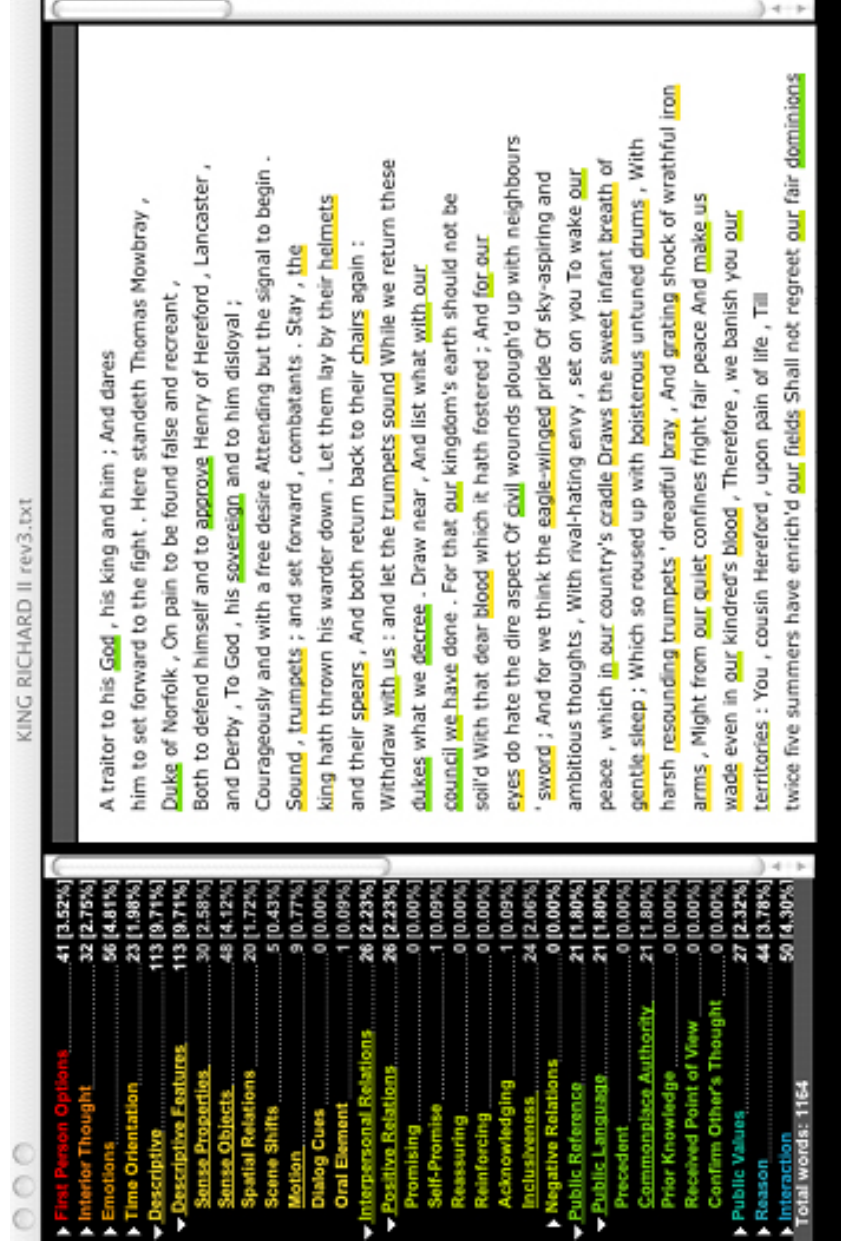


Figure 7: DocuScope screenshot illustrating exemplary History strings from *Richard II*, 1.3.



Figure 8: DocuScope screenshot illustrating exemplary History strings from *Romeo and Juliet*, 1.1.

ings is this: given that literary critical constructs are densely comparative and registered at potentially every level of our language, is a broader set of questions available to us now that we can use to study these interrelations at multiple levels of abstraction and conscious attention? The language of design might be helpful, particularly the notion of “affordances” or ranges of activity that are not precluded by the properties of a given material medium or arrangement of things.²⁵ What are the affordances of Shakespearean drama, and how are they registered or constrained by massively iterated linguistic activities, which we can track over an expansive range of texts in the growing corpus of digitized works? To what extent is linguistic filiation *merely* stylistic? Are there times when a writer, deciding to build a new sort of story on an old linguistic framework, registers something like a measurable cultural or ideological solidarity with past forms or attempts to stabilize emerging ones? These are not questions that can be answered simply by gathering data and counting things: they are fully interpretive, as is—we believe—all the work that goes under the name of algorithmic, digital, or iterative criticism.

Nor does this type of criticism invalidate previous forms of literary inquiry. If anything, it demonstrates that well-read, well-trained human beings are the most sensitive contraptions imaginable to differential phenomena like genre, and that this kind of judgment *is only approximated* (but provocatively so) by disaggregated tagging techniques coordinated by mathematical models. There is thus some basic similarity between the techniques that we are using, which call attention to exemplary patterns in the text (albeit patterns that have been statistically discerned), and the search for exemplarity that characterized the work of a great close reader like Erich Auerbach. Auerbach, often relying on memory as he was writing his landmark study of mimesis in Istanbul, toyed with the idea that the exemplary patterns he discerned in particular passages emerged from the occasional suspension of his own deliberate modes of attending to literary texts: “The great majority of the texts [for my study] were chosen at random, on the basis of accidental acquaintance and personal preference rather than in view of a definite purpose. Studies of this kind do not deal with laws but with trends and tendencies, which cross and complement one another in the most varied ways.”²⁶ Auerbach is saying something very powerful about the fluid nature of filiation among texts, a kind of filiation expressed in tendencies and crossings rather than “laws.” Crucially, his way into this world of variation was at least partly arbitrary: he was helped by the fact that his materials could not

²⁵ See James J. Gibson, “The Theory of Affordances,” in *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, ed. Robert Shaw and John Bransford (Hillsdale, NJ: Laurence Earlbaum, 1977), 67–82.

²⁶ Erich Auerbach, *Mimesis: The Representation of Reality in Western Literature*, trans. Willard R. Trask (Princeton: Princeton UP, 2003), 556.

be deliberately configured to support his intuitions. Iterative criticism incorporates this arbitrary feature in the form of a structured accident, where a mixture of deliberation and alienating distance characterizes our encounters with the text.²⁷ What differentiates such criticism from the analysis of tropes or semantic play within and across texts from a given period—for example, the more historically inflected work of Patricia Parker—is that the patterns sought are diffused so deeply into the built environment of the text that they cannot be attended to without some kind of (inhumanly) structured assistance.²⁸ But as we have seen from this study, those things that we *do* attend to as readers, and with great subtlety, are often connected to the linguistic rumble underneath.

The iterative study of Shakespeare and his genres offers us a new window into the study of complexity, framed in humanities learning. Texts are some of the most complicated multivariate objects in the world; genre is one stratum of that complexity. Indeed, there was an enormous amount of information transmitted in Heminges and Condell's simple decision to divide thirty-six of Shakespeare's plays into three groups. Given the complexity of these linguistic objects, the simple act of drawing circles around groups of plays speaks gigabytes about how and why the experience of drama can be one of "kinds." We have, in effect, reverse engineered some of the complexity of these kinds onto the page using DocuScope, but there is much more to be done. Shakespeare's plays, for example, need to be compared to those of his contemporaries, and the filiations of genres within and across different authors' works need to be understood. Just as important, we need to understand how genres change over time and when and how they accommodate variation and atypical diversions. And we need to better understand how to represent this information to an interested group of scholars needing to make reasonable comparisons among different results. In the end, reading diagrams may prove to be as difficult as reading texts: both are strategic redispositions of elements that can be experienced another way.²⁹

We close this essay, then, with a provocation. Figure 9 shows a detail of a dendrogram, illustrating a large body of Renaissance drama currently available in the Text Creation Partnership, tagged by DocuScope, and arrayed in terms

²⁷ On deliberate accidents and early modern notions of experimentation, see Michael Witmore, *Culture of Accidents: Unexpected Knowledges in Early Modern England* (Stanford: Stanford UP, 2001).

²⁸ See, for example, Patricia Parker's exemplary close readings of *Othello* and *Hamlet*, which trace a web of semantic and figurative correspondences between the plays and "larger discursive networks" that structure the language of privacy and accusation in "*Othello* and *Hamlet*: Dilation, Spying, and the 'Secret Place' of Woman," in *Shakespeare Reread: The Texts in New Contexts*, ed. Russ McDonald (Ithaca: Cornell UP, 1994), 105–46.

²⁹ On diagrammatic knowledge, see John Bender and Michael Marrinan, *The Culture of Diagram* (Stanford: Stanford UP, 2010).

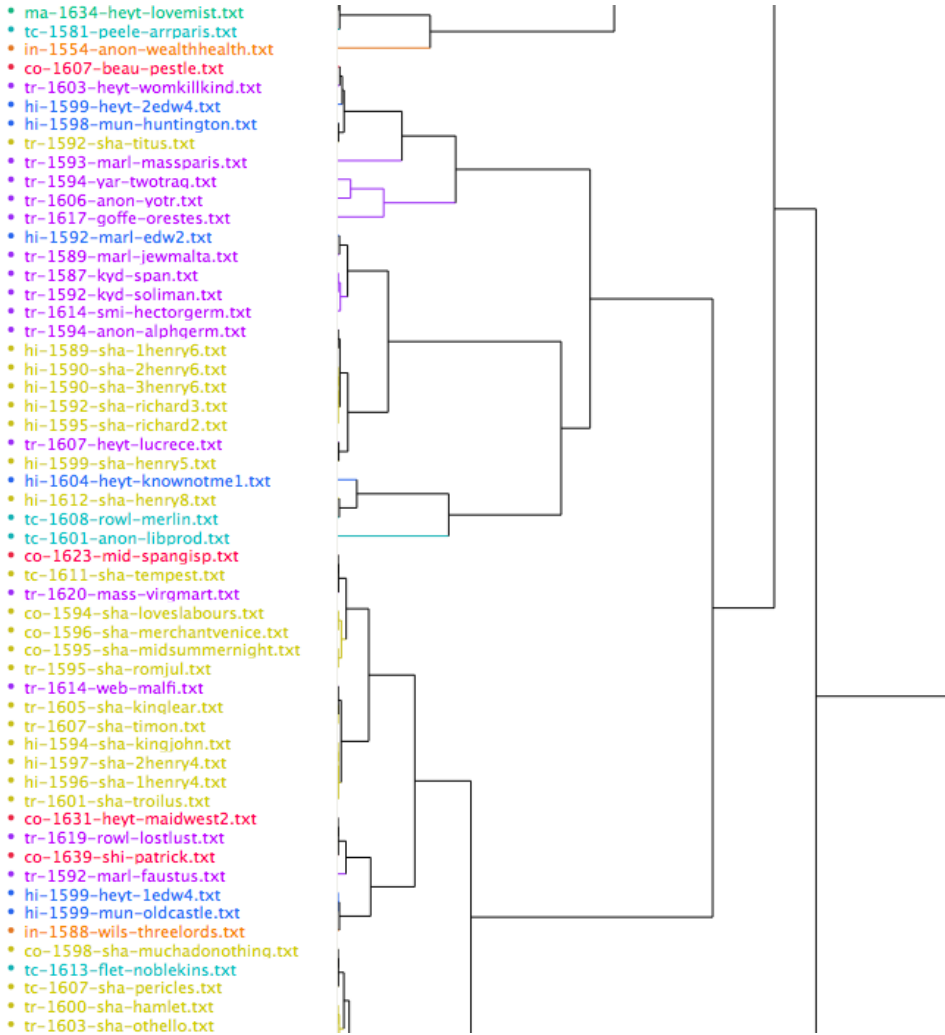


Figure 9: Section of a dendrogram (“The Very Large Dendrogram”) produced by Ward’s clustering method on scaled data using ninety-eight LATs to profile 320 plays written between 1519 and 1659.

of degrees of similarity.³⁰ Essentially, this is a diagram of linguistic similarity and difference approximately three hundred early modern theatrical texts: a snapshot of variation as it is patterned in a large group. We should stress that our findings here are provisional, part of what statistician John W. Tukey would call the exploratory phase of analysis.³¹ Certain caveats apply. The sample is not constructed to be representative; it was composed of texts available when we wrote this essay in early 2010. The metadata identifying genre, date, author, and title (which form the file names visible on the left-hand side of the diagram) are in some cases oversimplified or out of date. Finally, the statistical technique used to produce the dendrogram—cluster analysis by Ward's method—is known to produce broadly reliable overviews of data sets, but it necessarily simplifies the variation and relationships in a population.³²

Having offered these caveats, we nonetheless believe that it is important to codify foundational research questions and the available techniques for answering them in a pilot study such as this: these questions and techniques will be invaluable once we possess a complete sample of extant early modern drama.³³ Even in a pilot study, we can identify fascinating patterns for further study. To the question "What patterns variation between these plays?" we can answer, "A whole range of things." Looking at the bottom of the full dendrogram, for example, we see that almost all of Jonson's masques cluster together, while further up the diagram several of his comedies have clustered together (*Epicæne*, *Bartholomew Fair*, *The Alchemist*) alongside several comedies by Fletcher, suggest-

³⁰ We were lucky enough to get usable text files of these plays from Martin Mueller at Northwestern, who has developed some extremely powerful modernization procedures resulting in texts that are just as "countable" as those we studied in the hand-modernized Moby Shakespeare corpus. Mueller provisionally divided these plays up into generic groups using Alfred Harbage's *Annals of English Drama* and title page material. See Alfred Harbage, *Annals of English Drama 975–1700*, 3d ed., rev. S. Schoenbaum and Sylvia Stoler Wagonheim (London: Routledge, 1989). We have begun full-scale study of this corpus with Mueller in a joint research project between the University of Wisconsin–Madison, Strathclyde University, and Northwestern University. Mueller's work is documented at "DATA: Digitally Assisted Text Analysis," <http://literaryinformatics.northwestern.edu> (accessed 19 August 2010).

³¹ Interestingly enough, the humanities do not as a rule make explicit provision for the publication of exploratory scholarship, unless one argues that *all* products of reasoning in the humanities are offered with an unstated "as it were" that indicates the provisional nature of our assertions. Knowing how to silently add such a rider is a sure sign of membership in this research community. See John W. Tukey, *Exploratory Data Analysis* (Reading, MA: Addison-Wesley, 1977).

³² On the cluster analysis method developed by Joe H. Ward Jr., see H. Charles Romesburg, *Cluster Analysis for Researchers* ([Raleigh,] NC: Lulu.com, 2004), 134–35.

³³ Here, we harken back to John Unsworth's important 2005 work, "Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This?" at <http://www3.isrl.illinois.edu/~unsworth/Kings.5-00/primitives.html> (accessed 28 July 2010). We thank Alan Galey for calling our attention to this connection.

ing that sometimes genre—or even time of composition—trumps authorship as an organizing principle of similarity. Shakespeare's plays cluster in different groups, with the early history plays massing in one part of the diagram while the later history plays, along with some of the tragedies, clustering in another. *Two Noble Kinsmen* and *Pericles*, coauthored by Shakespeare, cluster with other Shakespeare plays (despite there being another hand in the drama). We can see evidence in this diagram for the linguistic effects of date, authorship, and genre. The work of disentangling these effects and making sense of the diagram is something only a highly trained human reader with a knowledge of the corpus of early modern drama can do. Indeed, such work resembles that of interpreting a text. The branches of the dendrogram point us toward linguistic similarities and differences which we might not have guessed at without this diagram; but the diagram does not tell us what those differences are, nor why they are important, nor why they exist in the population. Exactly what is a "population" of texts, and what are its natural or conventional temporal limits and its generic modes? To what extent do particular genres afford or allow generic deviation, and under what historical conditions? To what extent, finally, is *any* pattern of similarities among a group of texts simply a function of the population—generic, temporal, geographical—within which that difference becomes intelligible? A diagram like the one below may not itself provide the answers to such questions, but like any prosthetic, it points us toward better formulations of them and the provisional answers on which criticism thrives.