Stylometry
ISSN 2283-8759
pp. 42-74 (2020)

Computational Philology*

Jonathan P. Lamb

I begin with a walk down memory lane. The year was 1986, and critical theory had swept into most corners of literary studies. Yale theorist Paul De Man published a book called *Resistance to Theory*, about the emergence of literary theory in the American academy. De Man's defense of theory involved an unexpected but, as it turned out, highly influential claim, reflected in the title of the essay from which I want to quote: "The Return to Philology". De Man argued that "in practice, the turn to theory occurred as a return to philology, to an examination of the structure of language prior to the meaning it produces" (De Man 1986, 24, emphasis mine). For De Man and for many others in what has been called the "linguistic turn", at the heart of both critical theory and philology is a

* I offer enthusiastic thanks to Whitney Sperrazza, John Ladd, Doug Duhaime, Jonathan Hope, Laura Mielke, Anna Neill, Katie Conrad, Randy Fuller, Paul Outka, and Phil Drake. Thanks especially to Hugh Craig and the two anonymous reviewers for their formative feedback.

"concentrat[ion] on the way meaning is conveyed rather than on the meaning itself" (De Man 1986, 23).

In defending theory, De Man noted that "those who feel they may have to modify or to reconsider well-established pedagogical habits" come down with an "ill-humor". In the case of theory's detractors, the ill-humor runs deep:

> It feeds not only on civilized conservatism but on moral indignation. It speaks with an anxiety that is not only that of a disturbed tranquility but of a disturbed moral conscience. Nor is this mood confined to the opponents of theory. Its protagonists, in most cases, are just as nervous. When they appear not to be, their self-assurance often seems to be dependent on utopian schemes. (De Man 1986, 21)

In this "disturbed" context, De Man's identification of theory with philology largely succeeded. The reason we call the emergence of critical theory the "linguistic turn" (a phrase that must make actual linguists laugh) is that, for all its disciplinary genealogies, much critical theory of the late twentieth century addressed itself to the conditions of signification, stressing a homology between language and everything else. If everything works like language, the reasoning goes, the literary scholar rules.

Some things in literary criticism have changed since then. The hubris of critical theory lapsed some time ago, and new schools of critical discourse have arisen, including the use of computers to study the past[1]. In this new landscape, De Man's description of the resistance to theory may sound uncannily familiar to scholars working in computational text analysis. The often justified "ill-humor" of detractors and skeptics of computational methods has gone hand-in-hand with the increased use of those methods[2]. The "moral indignation" of many scholars toward computational methods – again, often warranted – is now at least a decade old, as is the nervousness among "protagonists" of these methods. Indeed, early theorizations of so-called distant reading took refuge in the

1 The ever-evolving *Debates in the Digital Humanities* series offers the best sense of the emergence of this field (Gold and Klein 2019).

2 The 2017 *PMLA* forum on distant readings offers a taste of the scholarly response. See especially Drucker 2017.

"utopian scheme" provided by the enhanced scale of computational methods[3].

This essay returns once again to philology to alleviate the apparent conflicts between conventional and computational methods. Although it may seem strange, even perverse, to invoke something so old-fashioned, so glacially slow as philology, my claim here is that computational text analysis – using computers to study literature – is basically philological. Digitally driven methods can, and in a certain sense must, extend the longstanding practices and techniques of philological scholarship, by way of, and not without reference to, the concerns of critical theory in its past and present forms. Given the vexed history of philology as the so-called origin of the humanities, the stakes are high (Turner 2014). By applying computational techniques to three different datasets (Shakespeare's plays, all early modern drama, and Shakespeare's characters), I offer a new handle for these methods: computational philology[4].

Philology and Computation

Although De Man's definition of philology was influential, it is hardly universal. The fact that many philologists dispute De Man's account of philology is precisely why I have quoted so heavily from it. I mention this point at the outset to emphasize that longstanding disagreements over philological scholarship reappear in disagreements over computational methods, and the questions surrounding philology are the same as those surrounding computational text analysis. The definition of philology itself is (or should be) relevant to what literary scholars make of computational methods (Watkins 1990). The broadest definition is this: philology is anything to do with the study of texts. James Turner defines it as "the multifaceted study of texts, languages, and the phenomenon of language itself" (Turner 2014, ix). Saussure defined it as the field of knowledge whose mission is "to correct, interpret, and comment

[3] See Ted Underwood's helpful pivot away from these utopian schemes (Underwood 2019, ix-xxii).

[4] Although many European scholars have used similar terms for some time, it has not caught on in English-language scholarship.

upon texts" (Ziolkowski 1990, 6). Roman Jakobson and Friedrich Nietzsche called philology "the art of reading slowly" (Lönnroth 2017, 15). More recently, Hans Gumbrecht defined it as "a configuration of scholarly skills that are geared toward historical text curatorship" (Gumbrecht 2003, 2). Admittedly, none of these jumbled definitions screams 'computers', which we tend to value for their ability to read quickly rather than slowly, to predict rather than comment, and to count rather than read.

Notwithstanding these rather indistinct definitions, philology clusters around a set of practices. This is what philology has involved, historically:

- Identifying texts
- Editing/curating texts
- Writing commentary
- Historicizing/contextualizing texts
- Tracing the histories of words, phrases, and forms
- Comparing instances of forms
- Principled generalization
- Identifying significant features
- Addressing iteration
- Describing/explaining textual variation[5]

These practices imply several key questions, which any appeal to philology must contend with: Is philology positivist? That is, does it strive to be empirical and scientific, and what would it mean if that were the case? How does philology relate to theory? Finally and most important, is philology *hermeneutic*? That is, does it make interpretive claims about texts?

Even those who want to answer 'yes' to this last question do not merely identify philology with interpretation. Rather, philology facilitates and even excites acts of interpretation by making texts available to analysis. The textual scholar David Greetham, for one, energetically rebuked the widespread belief that philology is "prehermeneutic" (i.e., before interpretation) on the grounds that the philological activities of textual studies and editing "can never

5 Even a list such as this may be controversial for its under- or over-representation of texts and language (Orlemanski 2015).

be prehermeneutical because [they are] already embedded, as cultural artifact[s], in the hermeneutic circle" (Greetham 1997, 19). In a gnarly but important sentence, Greetham argues that "just as criticism under poststructuralism became tropical and linguistic rather than extralinguistic and referential, so too did the operations of textual criticism become equally rhetorical and therefore just as 'threatening' to the imputed prehermeneuticism that de Man […] observe[s] for philology" (Greetham 1997, 16). In other words, of course philology is bound up with interpretation, because all of its activities belong within the hermeneutic circle. Even for Greetham, however, they remain separate components within that circle: the techniques of philology, themselves products of interpretative work, inevitably produce new interpretations, which in turn affect philological techniques.

Scholars have been addressing similar issues about computational text analysis for at least a decade. Can computers help us think things we have not thought before? How does computation fit with critical theory? Can an algorithm produce an interpretation? Should we expect it to? At the very least, these shared questions suggest a philological continuum on which computational study takes place. More tellingly, here is a list of activities common in computational study:

- Identifying texts
- Curating/regularizing texts
- Generating/manipulating markup
- Modeling contexts
- Tracing the histories of words, phrases, and forms
- Comparing instances of forms
- Principled generalization
- Identifying significant features
- Addressing iteration
- Describing/explaining textual variation[6]

My point here is to highlight the significant overlap between these two sets of practices – and not just overlap, but an isomorphism

6 A recent *PMLA* special issue puts many of these practices on display (Booth and Posner 2020).

between one, so established and foundational as to be invisible, and the other, still so emergent that people set their hair on fire when you start to talk about lexical distribution and predictive modeling.

Recent scholarship has brought these questions into sharp relief, infamously in Nan Z. Da's "The Computational Case against Computational Literary Studies". In forty pages of prose plus a GitHub appendix plus a Chronicle article plus multiple blog posts, Da accuses several scholars of misapplying statistical measures to the study of literature. For instance, Da reproduces a graph of Shakespeare's plays and writes that "you can't use [this approach] with the hope that [it] will work magic for you in producing interpretations that are *intentional*, that have meaning and insight defined with respect to the given field" (Da 2019, 621). Da, many of those against whom she writes, and others responding to the article seem to assume that interpretation is the point of computational text analysis. Da elsewhere assumes, in a puzzling contradiction, that computation does not imply a critical function[7].

But does everyone agree that computers *produce* interpretations? Of course, as Amelia Acker and Tanya Clement have reminded us, computation is hermeneutically loaded, in the sense that interpretative choices subtend every step of the computation process (Clement and Acker 2019). But that is very different from expecting a computer to make claims about significance. Scholars of computational stylistics and authorship attribution have long integrated this point into their various modes of inquiry, particularly with respect to studies based on word frequency counts (Burrows 1992; 2002; 2005; 2007; 2012; Jackson 1979; 2003; 2014; Craig 2004; Taylor and Egan 2017). Their explicit goals are rarely hermeneutic, even when they remain cognizant of the interpretive consequences of computational choices. Although I remain ambivalent toward much attribution scholarship, I take this virtue as salutary. I would suggest that if we view computational text analysis as an extension of philology, then the conflict among these approaches is not so sharp, because we have rightly adjusted our expectations.

7 The debate has continued to develop since Da's article appeared (Weatherby 2020; Underwood 2020; Da 2020).

Principal Components Analysis

Figure 1 shows a graph of Shakespeare's plays, similar to the one Da says cannot "produc[e] interpretations" (Da 2019, 621). This graph will look familiar to readers who have encountered similar graphs of early modern plays in the work of Jonathan Hope and Michael Witmore (2010), and more recently that of Hugh Craig and Brett Greatley-Hirsch (2017). To make this graph, I took the top 200 most frequent words in Shakespeare's plays (plus the anonymous comedy Mucedorus – more on that below) and ran the counts through a procedure called Principal Components Analysis, or PCA[8]. Craig and Greatley-Hirsch offer a helpfully accessible description of this technique:

> PCA […] is a statistical procedure used to explain as much of the total variation in a dataset with as few variables as possible. This [explanation] is accomplished by condensing multiple variables that are correlated with one another, but largely independent of others, into a small number of composite 'factors'. (Craig and Greatley-Hirsch 2017, 30-31)

PCA is thus a *dimension reduction* technique, in which we can identify a small number of components or factors – mathematical abstractions based on the features selected – responsible for a proportion of the dataset's variation. In PCA, the factor responsible for the greatest proportion of variation is usually called "principal component 1" or PC1. The factor responsible for the greatest proportion of the remaining variation is usually called "principal component 2" or PC2. The variation accounted for by this second factor must be uncorrelated to the variation accounted for by PC1; in other words, PC2 captures a different variation from that captured in PC1. Although we could calculate more principal

8 I used the Visualizing English Print texts of the plays, which represent transcribed versions of the First Folio texts ("Visualizing English Print" n.d.). To tokenize and normalize the data and create the PCA model, I used the SciKit-Learn package in Python, which uses a covariance matrix to decompose the data. For Python code and higher resolution images, see my GitHub repository: https://github.com/jonathanlamb. On PCA, see Alt 1990; Kachigan 1991; Jolliffe 2002; Konishi 2014; Gray 2017.

components (factors responsible for decreasing amounts of variation, uncorrelated to the previous components), identifying just the first two components allows us to plot those values in a two-dimensional space and regard those plots as representations of variation across the dataset.

Although many more extensive explanations exist in the scholarship, let me illustrate how PCA works by reference to the graph in Figure 1. Doing so will not only clarify the unavoidably knotty language in the preceding paragraph but will also begin to substantiate my claim that this form of analysis is basically philological. First, imagine a spreadsheet, in which there are 37 rows (the selected plays) and 200 columns (the 200 most common words in those plays – words such as "again", "thee", and "you"). Each cell in the spreadsheet contains a numerical value referring to the number of times a particular word appears in a particular play. Next, we must normalize the data across the dataset to account for the different lengths of each text. The text of *Hamlet* is roughly twice as long as *The Comedy of Errors*, so making statistical comparison based on raw counts of the 200 most common words would drastically skew results. Although there are several established ways to normalize data (and ways to avoid the problem altogether), I have calculated the relative frequency of each feature. The results of this procedure are illustrated in Figure 2. Each word in each text receives a weighted value based on the length of the play in which it appears. These new values allow terms to be more rigorously compared across different documents (plays) even when those documents are very different in length[9].

Next, we run the PCA. This procedure takes the normalized dimensional data (200 words across 37 plays) and reduces them to a smaller number of components, in this case two. As I said, these

---

[9] Admittedly, this method creates some problems even as it resolves others. A different method, used by Craig and Greatley-Hirsch (2017), is to split each text into chunks of predetermined length. A still different method is to normalize the data using TFIDF, which weights features based on their frequency in a single document and across the corpus. I experimented with these methods and found they lead to roughly similar (though of course not identical) results. The larger point, of course, is that these methods pose questions of philological consequence.

two dimensions represent a mathematical abstraction of the larger dataset, so it is difficult to state what they include. Nevertheless, the language of representation is apt: PC1 and PC2 *stand in for* the variation across these documents. It is as though we are telling the computer, by way of linear algebra, "draw a line through these plays that accounts for the greatest variation among them"; and then, "draw another, perpendicular line that accounts for the greatest remaining variation". In the PCA used to create the graph in Figure 1, PC1 accounts for 30.1% of variation in the data, and PC2 accounts for 17%. Each document (play) in our dataset receives a score in both PC1 and PC2 representing where that document stands in relation to the variation among all the documents. These scores can be positive and negative, such that a document can be positive or negative in PC1 and positive or negative in PC2, and we can graph those scores in Cartesian space.

In Figure 1, I have plotted PC1 on the X axis (horizontal), and PC2 on the Y axis (vertical). Here, for perhaps the first time in the process, the intuitive knowledge that comes with familiarity with Shakespeare's plays begins to interact with the mathematical representation of variation among them. In the terms I have established, for instance, *Much Ado About Nothing* scores high in PC1 (far right on the graph), while *King John* scores low in PC1 (far left), and both score about the same in PC2 (little vertical difference). This means that the variation represented in PC1 distinguishes *Much Ado* and *King John*, but with respect to the variation represented in PC2, the two plays are similar. By contrast, consider the placement of *Henry V*, *Macbeth*, and *The Two Gentlemen of Verona*. *Henry V* and *Macbeth* have similar values on PC1 and PC2 (thus they both appear in the top left), and the graph tells us that these two plays have values opposing *The Two Gentlemen of Verona* in PC1 and PC2. In terms of the variation expressed in PC1 and PC2, *The Two Gentlemen of Verona* is the opposite of *Henry V* and *Macbeth*. They are also, not fully coincidentally, very different kinds of play.

One virtue of PCA is its tendency to separate a set of documents into different kinds or clusters. "Kind" here could mean dramatic genre, or it could mean other, perhaps unexpected, categories. The crucial point is that there is an uncertain but suggestive relationship between, on the one hand, the lexical features underlying the statistical model and, on the other hand, the intuitive relationships and patterns that those models may call our attention to. In Figure 1, the histories tend to cluster on the left side of the graph, and the comedies on the right. PC1 reliably splits histories from comedies and even from tragedies, which hover around zero on PC1 (hence, not much horizontal movement). Those familiar with a similar graph from the work of Hope and Witmore will not be surprised to see that *Othello* has the linguistic profile of a comedy (Hope and Witmore 2010). This position reflects the comic register of much of the play's language, for example in the great temptation scene, which has both the features (i.e., the lexical patterns, visible to the computer) and the dramatic structures (i.e., formal categories, intuitive to a reader or viewer) of a wooing scene.

Other categorical relationships emerge on the graph, in many cases prompting us to look again at both the features informing the statistical procedure and at the plays themselves. The close proximity of *Timon of Athens* and *A Midsummer Night's Dream*, for example, reminds us that, in both plays, groups of characters go to the woods outside Athens. Again, the two plays' positions on the graph, which are a function of the PCA model built on the 200 most frequent words, remain distinct from the two plays' dramatic similarities. The computer does not know each play is set in Athens; strictly speaking, it does not know the two documents are plays at all. But the computational philologist can, and indeed must, reckon with both sets of observations. Likewise, the proximity of *Troilus and Cressida*, *Antony and Cleopatra*, *Cymbeline*, and *Julius Caesar* suggests a resemblance among classically oriented plays. The fact that they are classically oriented and the fact that they include a certain set of features producing their positions on the graph are distinct but perhaps meaningfully related. Moreover, I included the bestselling comedy *Mucedorus* in this dataset, based on my longstanding interest in the play as *like* a Shakespeare play though not written by Shakespeare (Lamb 2018). It clusters near *Romeo and Juliet*, a tragedy. The PCA model calls our preliminary attention to certain similarities between the two plays: both are about lovers who disguise themselves; both the male lovers have rivals and are exiled; both sets of lovers have their relationship threatened by the woman's angry father. Our awareness of this similarity arises from the linguistic profile modeled here, even if we must do much more work to build any kind of interpretation based on this resemblance.

But where is philology in all of this? That is a key question, since any exciting observations we might make about Figure 1 would seem to exist at many removes from the usual concerns of philology – language, text, textuality. I offer two important connections, one broad and one narrow, which I want to develop in the course of this essay. First, the 200 most frequent words used to build the PCA are, importantly, words, the meat and potatoes, as it were, of philological methods. Stylistics scholars have long viewed this particular class of lexical items (i.e., most frequent words) as especially important for studying style and register in a corpus. Even more broadly and historically, philology has addressed itself to *words in texts* as its primary objects of inquiry. Crucially, this emphasis on the compositional features of language carries across most modes of inquiry that regard themselves as philological[10].

Second, and more narrowly, this graph returns us to De Man's insistent claim that philology examines "the structure of language prior to the meaning it produces" (De Man 1986, 24). These computational techniques accomplish something similar, and similarly controversial. De Man's influential essay – entitled, you will recall, "The Return to Philology" – defended critical theory from its detractors by identifying it with philology. De Man championed the practice of close reading because it "cannot fail to respond to structures of language which it is the more or less secret aim of literary teaching to keep hidden" (De Man 1986, 24). Although Figure 1 engages in no conventional close reading, it accomplishes a similar response. I have just narrated a set of procedures that turn words into numbers, then into different numbers, then into mathematical abstractions, and then into graphical representations of those abstractions in two-dimensional space – illustrating Johanna Drucker's point that these are not so much "data" (given information) but "capta" (taken information) (Drucker 2011). These procedures draw attention to structures of language that, though not hidden by some 'secret' purpose, nevertheless become newly accessible (Froehlich 2020). If the graph

10 See, for instance, Masten 2016; Shore 2018. I eagerly await Ian Smith's in-progress work on "fair" as the marker of whiteness in *The Merchant of Venice*.

tells the truth, it is not about what Shakespeare's plays mean, but about the structures that make meaning possible and available to thought. In this respect, the methods employed here are philological in the narrow sense De Man offers.

To illustrate the layered philologies inherent in these methods, let me back up a step. Each play's scores in PC1 and PC2 are a composite of all 200 words in the dataset; the relative presence or absence of a single feature in a document affects that document's position on the graph. This affect is alternately known as a feature's "weighting" or "loading" (Craig and Greatley-Hirsch 2017, 35). Each feature affects the score a document receives on both PC1 and PC2; that is why we can't say what each principal component contains (i.e., some words rather than others), since all 200 features have weight in each component. In a move now commonplace in computational text analysis, after we 'fit' the PCA model to the 200 features and before we 'transform' that model to the 37 plays, we can extract the weighting scores for the 200 word features and graph them, as we did with the plays, in two-dimensional space. Figure 3 shows the features with the strongest weightings (greater than .05 in any direction); to graph all 200 would create unreadable clutter around the origin. Think of these points as vectors that 'push' plays around the graph: the weighting of "you", for instance, pushes texts that contain comparatively many instances of "you" higher in PC1, and somewhat higher in PC2. A glance at the text of *The Merry Wives of Windsor*, which appears on the top right in Figure 1, reminds us that it is a play in which an ensemble of characters relatively unfamiliar with one another constantly address each other with "you". We might even pursue this line of inquiry and read *Merry Wives* as thematizing the collapsing social class dimensions built into "you"/"thou" distinction. "King", meanwhile, weights texts lower in PC1 and not much at all in PC2; this loading likely has some relation to the position of history plays on the left in Figure 1. Finally, bear in mind that these vectors can cancel each other out in determining a given play's position. A closer look at the tragedies that are close to zero in PC1, for instance, may reveal high relative values of some left-pushing and some right-pushing features.

The point here is simply to emphasize the extent to which, as Andrew Piper has argued, "computational reading is inevitably tied to the norms and practices of the past" (Piper 2018, 3). To this statement, I would add that a method like PCA is tied not simply to the longstanding methodological norms of philology, but also to the controversies that accompany, contest, and critique those norms. Recall Greetham's insistence that De Man was wrong to suggest that philology is somehow "prehermeneutic", and that philological scholarship is always hermeneutically loaded. How comfortably the debate over computational methods falls along the continuum of disagreement over philology! Da's accusation that quantitative methods produce no interpretations, to which we will return, rehearses stages of the long debate over the status of philology.

Computational Philology at Scale

In his transformational book *Distant Horizons*, Ted Underwood argues that "[t]he point of distant reading is not to recover a complete archive of all published works but to understand the contrast between samples drawn from different periods or social contexts" (Underwood 2019, xx). Whereas earlier arguments on behalf of quantitative methods invoked textual recovery, Underwood suggests that modeling and dialectic make the most fruitful critical use of computational methods. Even here, we remain on philological ground, which has already known drastic increases of scale[11]. What "contrast between samples" emerges if we move from a single-author corpus to compare a much larger field, in this case a set of almost all the early modern plays available in Early English Books Online?

To produce Figures 4-6, I took similar steps as before, using the corpus of pre-1700 plays created by the Mellon-funded Visualizing English Print (VEP) project ("Visualizing English Print" n.d.). I added metadata to the corpus, then followed the same statistical procedure as before, using the 200 most frequent words across the whole corpus. Here, as with Shakespeare's plays, the very plays

[11] To pick the most obvious example, Erich Auerbach's *Mimesis: The Representation of Reality in Western Literature* takes as its focus not just "literature" but "reality"!

used constitute a series of selections that determine the results of the statistical procedures. This whole process of selection is a form of philological curatorship. Thinking about it this way – as connected to but not fully constitutive of interpretation – escapes the otherwise incisive accusations of Da and others. Computational methods are not for interpretation, even if they make interpretations possible. Rather, these tools are good for what philological work has always been good for: addressing what De Man calls "the way meaning is conveyed rather than the meaning itself" (De Man 1986, 23).

In Figure 4, which displays a PCA of all the dramatic texts in the VEP corpus labeled by genre, we once again see imperfect clusters. Masques and entertainments almost exclusively occupy the bottom right, while comedies dominate the bottom left. Just to repeat, this means that the two genres have different values in PC1 and similar values in PC2. Many tragedies have a positive PC2 value (thus appearing in the top half of the graph), but enough comedies appear in the same space that there is no clear, coherent distinction in terms of genre alone[12]. Likewise, most comedies have a negative value in PC1 (thus appearing on the left half of the graph), but again the distinction remains murky. Figure 5, in which dramatic texts have now been labeled with the decade in which they were written, helps clarify the situation. The comedies on the bottom left tend to have later dates (mid to late seventeenth century), while the comedies above the X axis tend to have earlier dates (late sixteenth and early seventeenth centuries)[13]. Relabeling the data by author – another choice with hermeneutically loaded assumptions, given what we know about collaboration and authorship in early modern dramatic texts – suggests that the PCA implemented here does not capture meaningful categorical distinctions of authorship, though some author clusters appear (see below). Here again, however, these observations create more questions than they settle. Is the

---

12 The ratio of variance explained in the model is relevant here: PC1 captures 37.1% of the variation in the dataset, while PC2 captures 8.1%. These numbers mean that horizontal position on the graph represents a substantial proportion of the variation, while vertical position represents appreciably less, though still a significant amount of, variation.

13 PC2 and date of writing have a negative Pearson correlation of -.495 ($p = 1.2^{-26}$). That is, the later the date of writing, the lower the PC2 value.

variance captured in PC2, which seems to have some imperfect relation to date of writing, a function of the changes to the English language in the seventeenth century? Is the variance captured in PC1 largely a function of genre, or some other category that also affects genre?

Like the feature selections and transformations that permitted it to be built, this descriptive model is best understood as a philological artifact. In this line of thinking, the virtue of the graph lies not in its explanatory power but in its philological potential to suggest, reframe, and motivate new critical work. Scholars of digital methods have been making this point for a long time, though not by reference to philology[14]. The fact that the graph shows us some things we already knew guarantees its coherence, and, at the same time, it propels us to new forms of inquiry. Just to choose one example, it may not surprise us to see that English translations of Seneca's plays appear near each other (see Figure 6). But precisely because that cluster does not surprise us, we must reckon with the fact that near this cluster also appear several plays by Christopher Marlowe (*1* and *2 Tamburlaine* and *Dido, Queen of Carthage*). Moreover, this cluster also includes Fletcher's *The Faithful Shepherdess*, Milton's *Samson Agonistes*, and Shakespeare's *Richard II* – an unlikely grouping if there ever was one. Even before we explore the features that have produced this cluster, philology furnishes an anthropology of language with which to approach these renderings of mathematical abstraction. Such an approach might begin with the observation that each of these plays concerns, in its dramatic action, failure. More precisely, these plays feature characters who achieve dramatic personhood *by way* of failure – the essence of the paradoxical stoicism that Seneca has been taken to represent (Lamb 2017). The next (philo)logical step would be to explore how the lexical features that produced this cluster relate to these thematic observations.

Following Underwood's point about comparisons across samples, this second PCA allows helpful contextualization of the first PCA of Shakespeare's plays. Making such a comparison is like a philologist contextualizing one writer's use of a word by reference to other writers' uses. Nay, it *is* – I know not "like". While *Othello*

[14] This point is on full display in Gold and Klein 2019.

appeared with the comedies in the first PCA, in this second model it falls near other tragedies (see Figure 7). Perhaps this seems unexciting, until we see the tragedies nearby: *The Miseries of Enforced Marriage*; *Darius, King of Persia*; *The Tragedy of Mustapha*; *Aureng-Zebe*; *Love's Sacrifice*. What we lose in productive generic confusion we gain with a whole new dramatic context for this play. These plays concern love, marriage, and the dissolution of both; they concern racialized outsiders from Persia and the Ottoman Empire; and, crucially for our approach to *Othello*, they dramatize conflicts among love, marriage, and race. These resemblances do not immediately explain the word counts that underlie this graph, however, and that is precisely the point: the model, based on a series of selections and abstractions, produces new cues for scholarship. It is unhelpful to call this process EDA, or Exploratory Data Analysis (Da 2020), because that label cuts off the deep philological roots of the process.

In articulating the value of computational methods, Piper complains that conventional literary studies lack two important things: first, a way to address how much recurrence happens in literature – how many repeated genres, modes, forms, and so on there are, and second, a science of generalization – the ability to move from a single instance to a more general claim. He rightly claims that computation can provide these two things (Piper 2018, ix-xi). With all due respect to Piper, however, we already had a set of techniques for doing both of these things. It's called *philology*. Understood as standalone, interpretative expressions of information, these graphs might as well have the minimal value attributed to them by critics. But understood as reagents in the rich anthropology of language we call philology, they become much less ambitious, perhaps, but also far more exciting.

Every Man in His Humour

The literary theorist Barbara Johnson asked a great question: "[w]hat if the philologist's attentiveness to language were great enough to open up irresolvable difficulties, *resistances* to meaning, or other, unexpected meanings within the text?" (Johnson 1990, 28). "What is at stake" in philology, she wrote, is "how to read in such a way as to break through preconceived notions of meaning in order to encounter unexpected otherness" (29). That project, I want to suggest in this final section, continues in the work of computational philology. Many scholars have written eloquently about the genuine problems of computational techniques, which are embedded in histories of racism and misogyny (McPherson 2012; Gallon 2016). In step with this critique, I worry too over the nationalist history of philology (Richardson 1994; Benes 2019). Engaging in both fields requires profound and continual self-critique, which in turn gives a sharp edge to the hermeneutic possibilities and futures of these methods.

I offer one more PCA, this one moving beyond whole plays as the principle of selection. This time, I collected the speech of Shakespeare's characters who speak more than 500 words and ran a PCA model with those data (see Figure 8)[15]. Again, there's a vaguely discernible generic break between history (left) and comedy (right). But what really excites me here is the way the characters break down imperfectly into quadrants. In the top left, we have medieval and classical characters: Hotspur, Henry V, Ulysses, and Agamemnon. In the bottom left, we find characters who tell people what to do, whether or not those people listen: the Ghost in *Hamlet*, John of Gaunt, Queen Margaret (twice), Tamora and Saturninus, and the Fool in *King Lear*. In the bottom right, we find many women characters: Miranda, Cordelia, Hermia and Helena, Isabella, Katherine (from *The Taming of the Shrew*), Desdemona. Finally, in the top right, we find servants and clowns: Camillo, Rosencrantz, Nerissa, Dogberry. We also have the Duke in *Measure for Measure* and Rosalind, who spend most of their plays performing lowliness.

---

15 Although many of the most talkative characters tend to appear closer to the middle of the graph, there is no automatic strong correlation. The Pearson correlation between the absolute value of PC1 and length of part is -.005 (p = .917), not significant, whereas the correlation between the absolute value of PC2 and length of part is -.135 (p = .008), highly significant. I would like to acknowledge the work of Michael Poston, who created the Folger Digital Texts API (https://www.folgerdigitaltexts.org/api) that makes this curation quite simple. I also thank the incomparable Nellie Kassebaum for her collaborative efforts on behalf of this project.

One useful effect of this graph is to decompose the findings of the first two PCAs. A dramatic text is made up primarily of character voices, and those characters represent individuated persons speaking in distinctive mimetic patterns and rhetorical situations – almost like real people. *Othello*'s position on the graph is largely a function of the words of its most talkative characters; studying where its characters fall in the character speech corpus helps contextualize previous findings. In Figure 9, perhaps to our surprise, Othello and Brabantio appear in nearly the same spot (right near Titus, Lear, and Cleopatra – a take-charge group if ever there was one). Roderigo, Desdemona, and Emilia appear close together as well, near four major characters from *Twelfth Night* (Malvolio, Sebastian, Olivia, and Viola). These seven characters have in common a sustained interactivity with other characters, including each other. Iago, meanwhile, appears near no other *Othello* characters, but he does share a speech profile with *Much Ado*'s Don Pedro and *Hamlet*'s Polonius! This trio spends much of their stage time giving advice to others, including characters already mentioned. These observations may not mathematically explain why *Othello* appears with the comedies in Figure 1, but they do manifest dramatic factors underlying those mathematical representations.

As before, we can extract the weightings of the 200 features used to generate Figure 10 and use them to formulate critical questions. This well-established rhetorical move in the digital humanities often presents itself as a departure from conventional literary scholarship methods, and from more recent forms of cultural critique. In a philological context, however, this is a familiar move, interrogating how the observed state of a text came to be that way, and how it changed over time. In Figure 10, I have graphed the features with the highest loadings (greater than .02); their position on the graph represents their loading vectors in PC1 and PC2. As with the feature graph of Shakespeare's plays, personal pronouns strongly affect a character's position. "You" and "your", the pronouns used by social inferiors, pull characters toward the bottom-right quadrant. By contrast, "thee", "thy", and "thou" pull toward the top-left, where we find bossy characters. Looking at Shakespeare's characters this way reminds us of what they always were: clusters of signifiers arranged in such a way that we infer something beneath them – a body beneath the clothes. As I indicated, scholars trained in philological methods are pretty good at collating textual variants at various levels of abstraction. This graph too invites us to collate clusters of signifiers by way of abstraction – for example, Juliet and Caliban, who appear in almost exactly the same position on the graph (see Figure 8). A kind of statistical Hinman Collator.

Together, these graphs invite us to regard a given play as a composite in two related and perhaps competing senses. First, the play is composed of artfully selected and disposed lexical items across the entire text, and second, the play is a composite of individuated characters speaking a more peculiar set of lexical items in particular situations. Plays are certainly more than that, but they are not less. When we encounter what Johnson calls "unexpected otherness" (Johnson 1990, 29), as for example in the resemblance between Juliet and Caliban, we need a critically self-aware philology. Figure 10 suggests – and a look at their speeches could confirm – that these two characters appear where they do partly because they use "thou" pronouns along with "me" and "my". Both characters speak from positions of abject familiarity, though their abjections are very different. Both are forbidden by an authority figure (Capulet, Prospero) from pursuing different kinds of erotic attachments with another character (Romeo, Miranda). Both respond to advice from a character perceived as knowledgeable (Father Lawrence, Stephano and Trinculo). Both are disempowered when political changes in which they have some stake displace their personal ambitions (marriage, insurrection). Drawing on decades of postcolonial readings of Caliban, we might ask, do the differences within these characters' similarities make it possible to interpret Juliet as a figure marked by whiteness?

To be clear, none of these comparisons is *in* the data. Many are admittedly fanciful. But all are energized and perhaps inspired by the data, and that is precisely the value of thinking about these graphs philologically. If we lower our expectations about what these methods are supposed to produce from a fully-formed, robust hermeneutics to a philologically-oriented exploration of language, then we may find that we can raise our ambitions about the intellectual work they make possible. Circling back to De Man, these methods seek to address the language of texts as objects of analysis at varying levels of complexity and abstraction. Moreover, they involve curation at every step – curation of documents in new or different configurations, curation of texts in new patterns, curation of linguistic models. They also enable us to abstract and therefore compare texts on the way to making an interpretive claim. These techniques do not and can never constitute interpretation (just as nothing philology does is merely an interpretive act); rather, they excite interpretation and make it possible.

Readers may have noticed that I have not quoted from any primary texts in this essay. This is by design. The urge to move past computational methods remains so strong in literary studies, even among those attracted to these methods, that I wanted to dwell on the models as philological composites in their own right. The curation choices, statistical procedures, and graphical representations offered here are not "prehermeneutic", in Greetham's terms. They do not exist in some neutral, apolitical zone, just as the philological work that underlies most literary scholarship does not. We need a scholarship capacious enough to include statistical models alongside textual variants, semantic contexts, theories of textuality, and cultural critique. We need, in short, a computational philology.

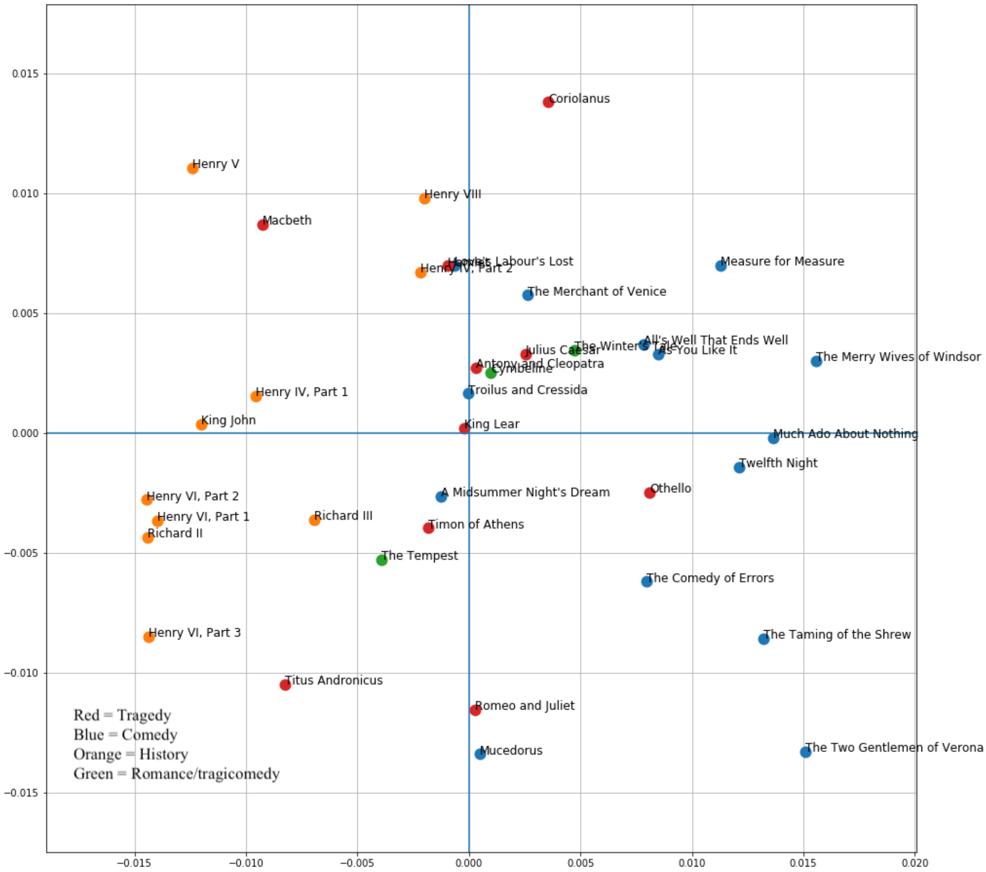


Fig. 1

	again	against	all	am	an	and	any	are	art	as	...	whose	why	will	with	word
0	0.000839	0.000789	0.004538	0.003650	0.001726	0.031469	0.001529	0.005376	0.001381	0.007991	...	0.000296	0.002811	0.008089	0.009470	0.000789
1	0.000700	0.000088	0.004816	0.001926	0.002277	0.027581	0.001313	0.002539	0.001138	0.005954	...	0.001401	0.004728	0.007880	0.010419	0.000263
2	0.001427	0.000649	0.005903	0.002595	0.001622	0.029448	0.001297	0.004800	0.001881	0.007265	...	0.001038	0.000908	0.005384	0.008562	0.000584
3	0.001027	0.000540	0.002648	0.004702	0.002162	0.025943	0.001675	0.004648	0.001459	0.007945	...	0.000378	0.002378	0.008756	0.008972	0.000649
4	0.001193	0.001325	0.007023	0.002915	0.002032	0.029283	0.000883	0.006316	0.000265	0.006934	...	0.000574	0.000530	0.004461	0.007685	0.000133

Fig. 2

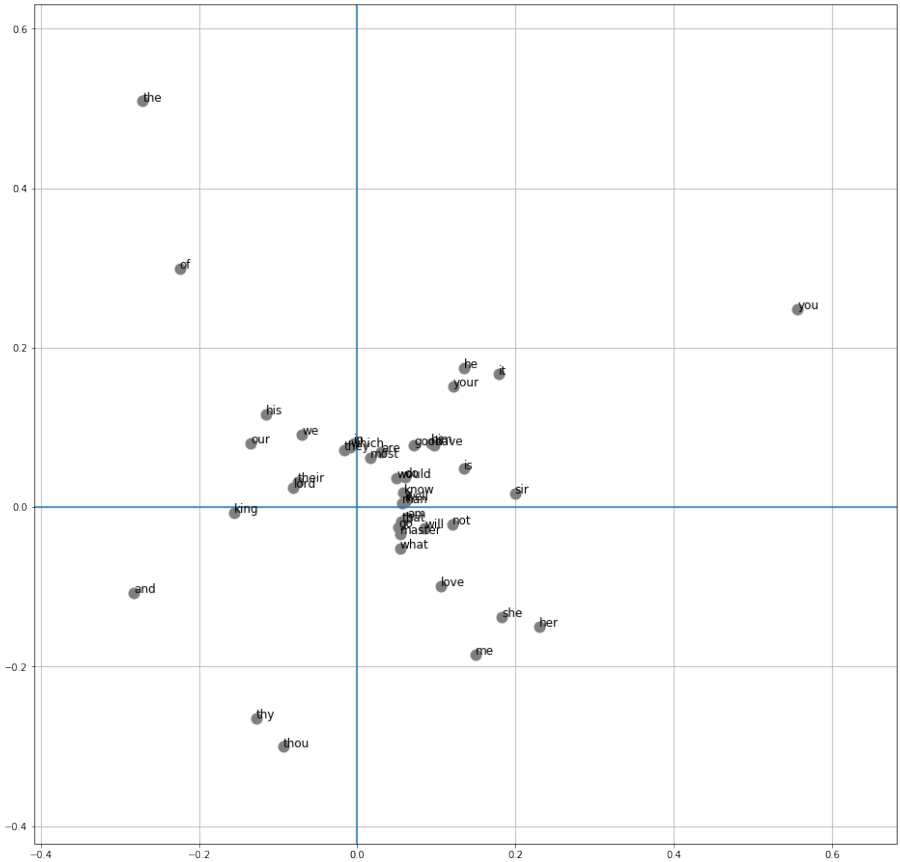

0.6
0.4
0.2
0.0
−0.2
−0.4
−0.4
−0.2
0.0
0.2
0.4
0.6
the
of
you
he
it
your
his
we
our
they
which
are
most
good
is
sir
their
lord
king
not
will
master
what
love
she
her
me
and
thy
thou

Fig. 3

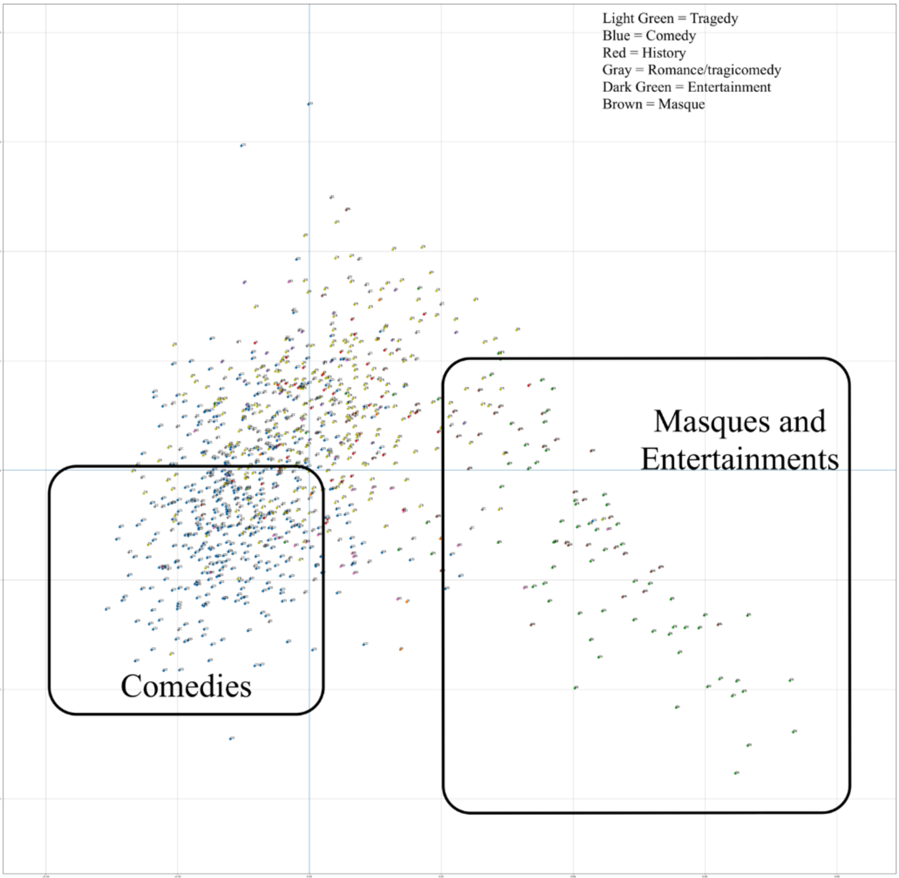

Light Green = Tragedy
Blue = Comedy
Red = History
Gray = Romance/tragicomedy
Dark Green = Entertainment
Brown = Masque
Masques and Entertainments
Comedies

Fig. 4

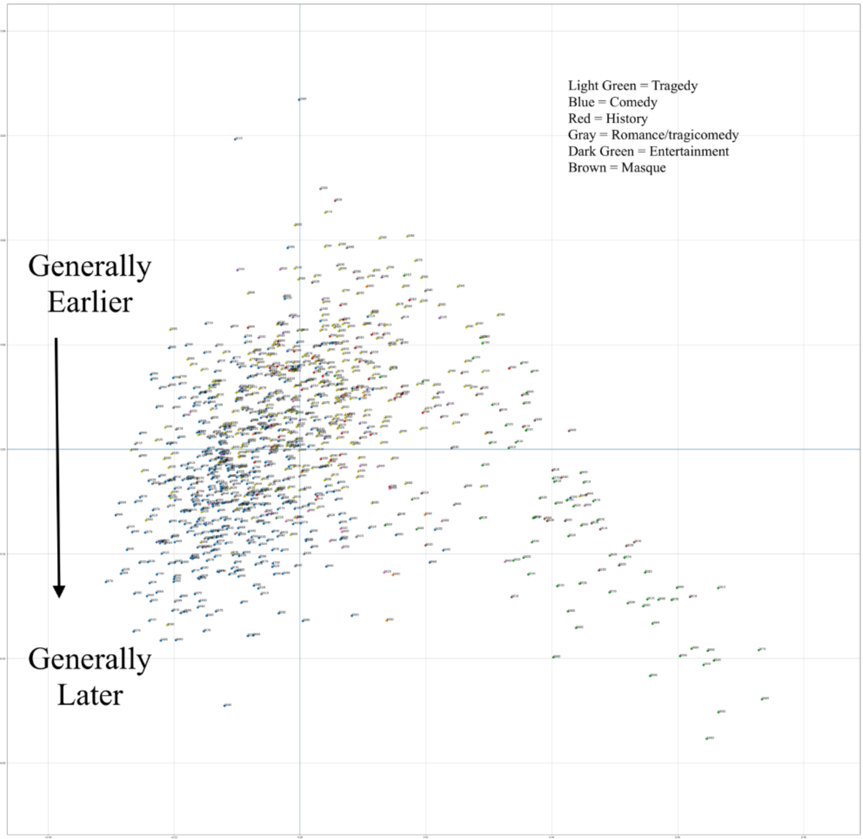

Light Green = Tragedy
Blue = Comedy
Red = History
Gray = Romance/tragicomedy
Dark Green = Entertainment
Brown = Masque
Generally Earlier
Generally Later

Fig. 5

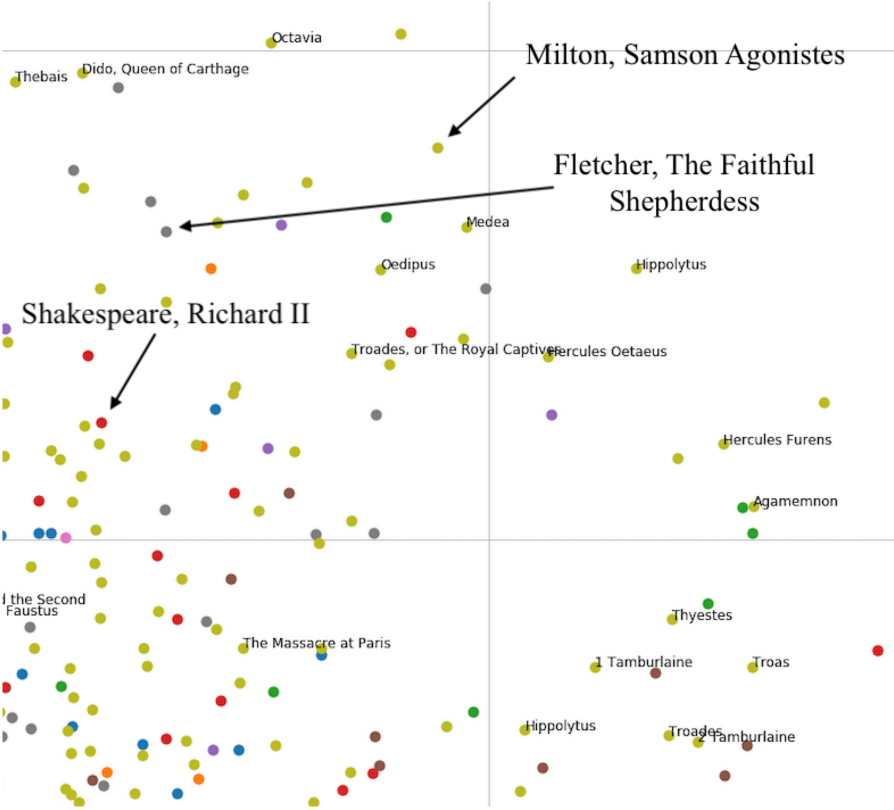
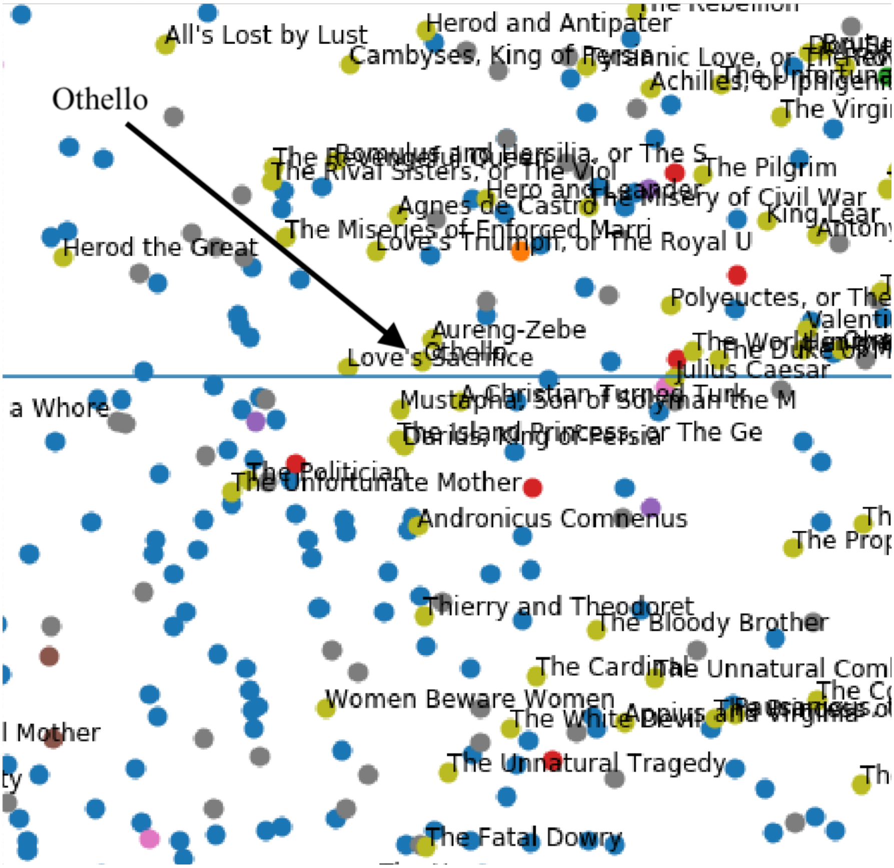
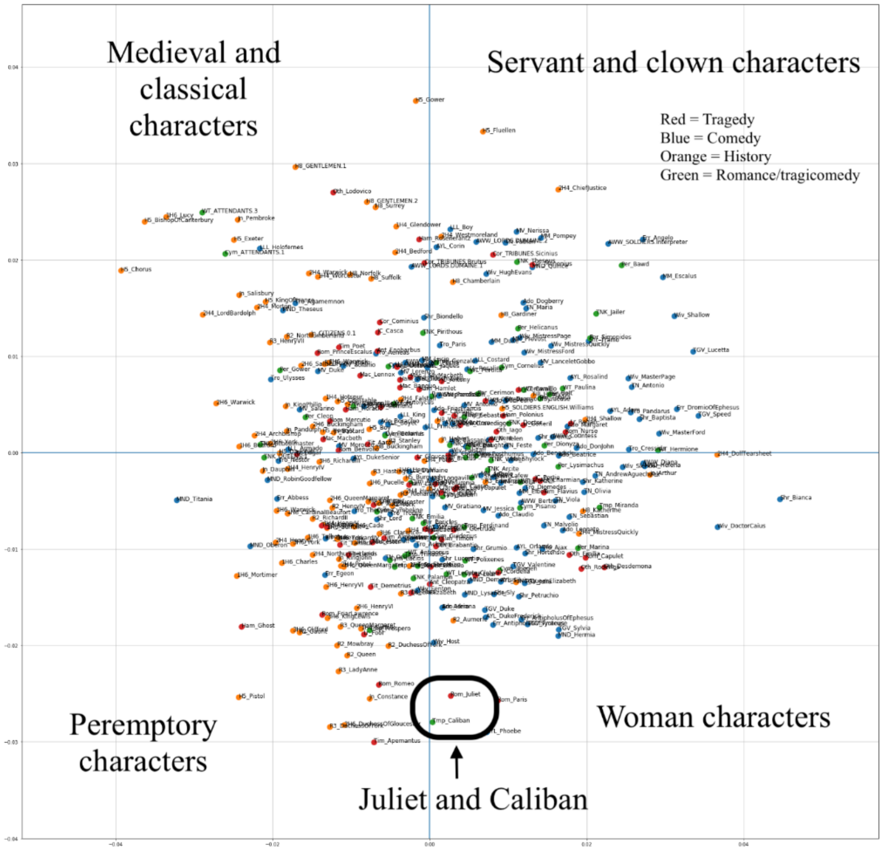

Octavia
Milton, Samson Agonistes
Thebais
Dido, Queen of Carthage
Fletcher, The Faithful Shepherdess
Medea
Oedipus
Hippolytus
Shakespeare, Richard II
Troades, or The Royal Captives
Hercules Oetaeus
Hercules Furens
Agamemnon
the Second
Faustus
Thyestes
The Massacre at Paris
1 Tamburlaine
Troas
Hippolytus
Troades
2 Tamburlaine

Fig. 6


Othello
All's Lost by Lust
Herod and Antipater
Cambyses, King of Persia
Achilles, or Iphigenia
The Virgin
The Rival Sisters, or The Viol
Romulus and Hersilia, or The S
The Pilgrim
Hero and Leander
The Misery of Civil War
Agnes de Castro
King Lear
Herod the Great
The Miseries of Enforced Marri
Love's Triumph, or The Royal U
Polyeuctes, or The
Aureng-Zebe
Love's Sacrifice
Othello
Julius Caesar
A Christian Turned Turk
Mustapha, Son of Solyman the M
The Island Princess, or The Ge
Darius, King of Persia
a Whore
The Politician
The Unfortunate Mother
Andronicus Comnenus
Thierry and Theodoret
The Bloody Brother
The Cardinal
Women Beware Women
The White Devil
Appius and Virginia
Mother
The Unnatural Tragedy
The Fatal Dowry
Fig. 7



Medieval and classical characters
Servant and clown characters
Red = Tragedy
Blue = Comedy
Orange = History
Green = Romance/tragicomedy
Peremptory characters
Woman characters
Juliet and Caliban

Fig. 8

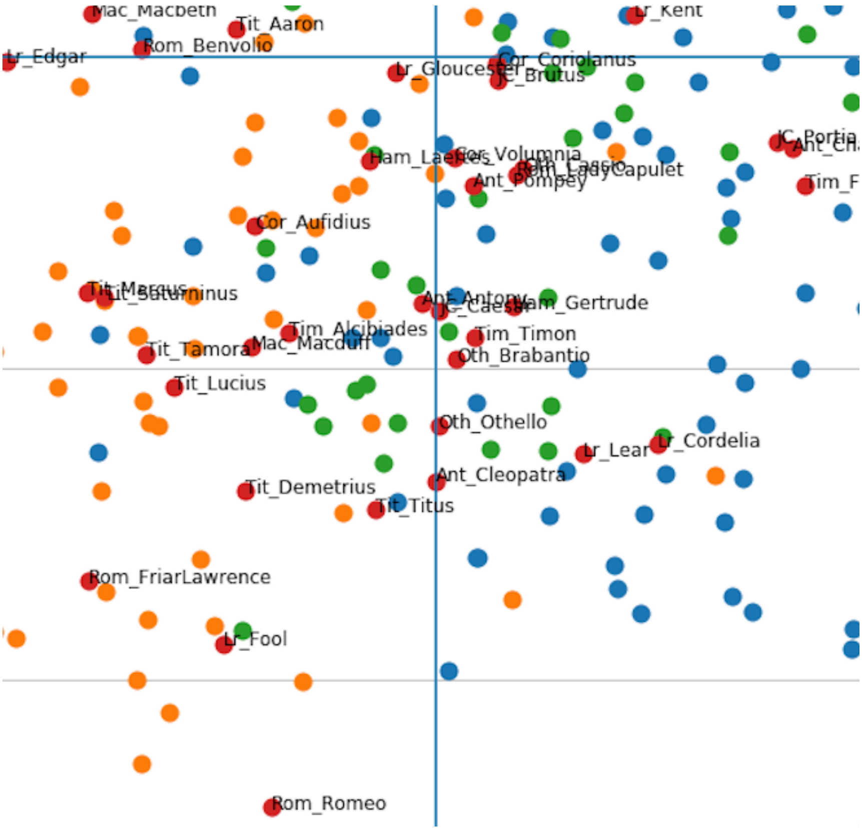

Mac_Macbeth
Tit_Aaron
Lr_Kent
Rom_Benvolio
Lr_Edgar
Lr_Gloucester
Cor_Coriolanus
JC_Brutus
JC_Portia
Ham_Laertes
Cor_Volumnia
Oth_Cassio
Rom_LadyCapulet
Ant_Pompey
Cor_Aufidius
Ham_Gertrude
Tim_Alcibiades
Tim_Timon
Tit_Tamora
Mac_Macduff
Oth_Brabantio
Tit_Lucius
Oth_Othello
Lr_Lear
Lr_Cordelia
Ant_Cleopatra
Tit_Demetrius
Tit_Titus
Rom_FriarLawrence
Lr_Fool
Rom_Romeo

Fig. 9

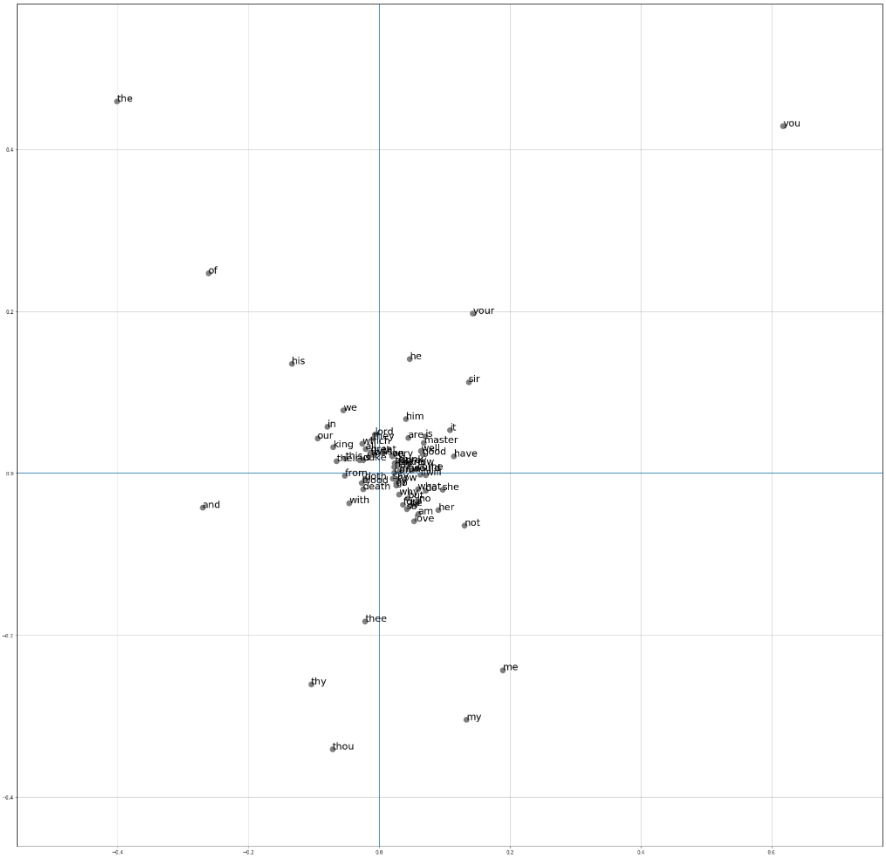


Fig. 10

References

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

Benes, Tuska. 2019. "Philology, Language, and the Constitution of Meaning and Human Communities." In *The Cambridge History of Modern European Thought*, edited by Peter E. Gordon and Warren Breckman, vol. 1 of 2, 330-52. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316160855.015.

Booth, Alison, and Miriam Posner, eds. 2020. "Special Topic: Varieties of Digital Humanities." *PMLA* 135, no. 1 (January).

Burrows, John. 1992. "Computers and the Study of Literature." In *Computers and Written Texts*, edited by Christopher S. Butler, 167-204. Oxford: Blackwell.

Burrows, John. 2002. "The Englishing of Juvenal: Computational Stylistics and Translated Texts." *Style* 36, no. 4 (Winter): 677-98. https://www.jstor.org/stable/10.5325/style.36.4.677.

Burrows, John. 2005. "Numbering the Streaks of the Tulip? Reflections on a Challenge to the Use of Statistical Methods in Computational Stylistics." *Digital Studies/Le Champ Numérique*, no. 1 (August). https://doi.org/10.16995/dscn.230.

Burrows, John. 2007. "All the Way Through: Testing for Authorship in Different Frequency Strata." *Literary and Linguistic Computing* 22, no. 1 (April): 27-47. https://doi.org/10.1093/llc/fqi067.

Burrows, John. 2012. "A Second Opinion on 'Shakespeare and Authorship Studies in the Twenty-First Century'." *Shakespeare Quarterly* 63, no. 3 (Fall): 355-92. https://doi.org/10.1353/shq.2012.0038.

Clement, Tanya, and Amelia Acker. 2019. "Data Cultures, Culture as Data: Special Issue of Cultural Analytics." *Journal of Cultural Analytics* (April). https://doi.org/10.22148/16.035.

Craig, Hugh. 2004. "Stylistic Analysis and Authorship Studies." In *A Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 273-88. Malden, MA: Blackwell.

Craig, Hugh, and Brett Greatley-Hirsch. 2017. *Style, Computers, and Early Modern Drama: Beyond Authorship*. Cambridge: Cambridge University Press.

Da, Nan Z. 2019. "The Computational Case against Computational Literary Studies." *Critical Inquiry* 45, no. 3 (Spring): 601-39. https://doi.org/10.1086/702594.

Da, Nan Z. 2020. "*Critical Response III*. On EDA, Complexity, and Redundancy: A Response to Underwood and Weatherby." *Critical Inquiry* 46, no. 4 (Summer): 91324. https://doi.org/10.1086/709230.

De Man, Paul. 1986. *The Resistance to Theory*. Minneapolis: University of Minnesota Press.

Drucker, Johanna. 2011. "Humanities Approaches to Graphical Display." *Digital Humanities Quarterly* 5, no. 1.

http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html.

Drucker, Johanna. 2017. "Why Distant Reading Isn't." *PMLA* 132, no. 3 (May): 628-35. https://doi.org/10.1632/pmla.2017.132.3.628.

Froehlich, Heather. 2020. "Dramatic Structure and Social Status in Shakespeare's Plays." *Journal of Cultural Analytics* (April). https://doi.org/10.22148/001c.12556.

Gallon, Kim. 2016. "Making a Case for the Black Digital Humanities." In *Debates in the Digital Humanities 2016*, edited by Matthew K. Gold and Lauren F. Klein, 42-49. Minneapolis: University of Minnesota Press. https://dhdebates.gc.cuny.edu/projects/debates-in-the-digital-humanities-2016.

Gold, Matthew K., and Lauren F. Klein, eds. 2019. *Debates in the Digital Humanities 2019*. Minneapolis: University of Minnesota Press. https://dhdebates.gc.cuny.edu/projects/debates-in-the-digital-humanities-2019.

Gray, Virginia, ed. 2017. *Principal Component Analysis: Methods, Applications and Technology*. New York: Nova Science Publishers.

Greetham, D. C., ed. 1997. *The Margins of the Text*. Ann Arbor, MI: University of Michigan Press.

Gumbrecht, Hans Ulrich. 2003. *The Powers of Philology: Dynamics of Textual Scholarship*. Urbana: University of Illinois Press.

Hope, Jonathan, and Michael Witmore. 2010. "The Hundredth Psalm to the Tune of 'Green Sleeves': Digital Approaches to Shakespeare's Language of Genre." *Shakespeare Quarterly* 61, no. 3 (Fall): 357-90. https://doi.org/10.1353/shq.2010.0002.

Jackson, MacDonald P. 1979. *Studies in Attribution, Middleton and Shakespeare*. Salzburg: Institut für Anglistik und Amerikanistik, Universität Salzburg.

Jackson, MacDonald P. 2003. *Defining Shakespeare: "Pericles" as Test Case*. Oxford: Oxford University Press.

Jackson, MacDonald P. 2014. *Determining the Shakespeare Canon: "Arden of Faversham" and "A Lover's Complaint."* Oxford: Oxford University Press.

Johnson, Barbara. 1990. "Philology: What Is at Stake?" *Comparative Literature Studies* 27, no. 1: 26-30. https://www.jstor.org/stable/40246725.

Jolliffe, I. T. 2002 [1986]. *Principal Component Analysis*. 2nd ed. New York: Springer-Verlag.

Kachigan, Sam Kash. 1991 [1982]. *Multivariate Statistical Analysis: A Conceptual Introduction*. 2nd ed. New York: Radius Press.

Konishi, Sadanori. 2014. *Introduction to Multivariate Analysis: Linear and Nonlinear Modeling*. Boca Raton, FL: CRC Press.

Lamb, Jonathan P. 2017. *Shakespeare in the Marketplace of Words*. Cambridge: Cambridge University Press.

Lamb, Jonathan P. 2018. "William Shakespeare's *Mucedorus* and the Market of Forms." *Renaissance Drama* 46, no. 1 (Spring): 57-86. https://doi.org/10.1086/697174.

Lönnroth, Harry, ed. 2017. *Philology Matters! Essays on the Art of Reading Slowly*. Leiden: Brill.

Masten, Jeffrey. 2016. *Queer Philologies: Sex, Language, and Affect in Shakespeare's Time*. Philadelphia: University of Pennsylvania Press.

McPherson, Tara. 2012. "Why Are the Digital Humanities So White? Or Thinking the Histories of Race and Computation." In *Debates in the Digital Humanities*, edited by Matthew K. Gold, 139-60. Minneapolis: University of Minnesota Press. https://dhdebates.gc.cuny.edu/projects/debates-in-the-digital-humanities.

Orlemanski, Julie. 2015. "Philology and the Turn Away from the Linguistic Turn." *Florilegium* 32 (January): 157-81. https://doi.org/10.3138/flor.32.007.

Piper, Andrew. 2018. *Enumerations: Data and Literary Study*. Chicago: The University of Chicago Press.

Richardson, Peter. 1994. "The Consolation of Philology." *Modern Philology* 92, no. 1 (August): 1-13. https://www.jstor.org/stable/438220.

Shore, Daniel. 2018. *Cyberformalism: Histories of Linguistic Forms in the Digital Archive*. Baltimore: Johns Hopkins University Press.

Taylor, Gary, and Gabriel Egan, eds. 2017. *The New Oxford Shakespeare: Authorship Companion*. Oxford: Oxford University Press.

Turner, James. 2014. *Philology: The Forgotten Origins of the Modern Humanities*. Princeton: Princeton University Press.

Underwood, Ted. 2019. *Distant Horizons: Digital Evidence and Literary Change*. Chicago: The University of Chicago Press.

Underwood, Ted. 2020. "*Critical Response II*. The Theoretical Divide Driving Debates about Computation." *Critical Inquiry* 46, no. 4 (Summer): 900-12. https://doi.org/10.1086/709229.

"Visualizing English Print." n.d. Accessed May 31, 2017. http://graphics.cs.wisc.edu/WP/vep.

Watkins, Calvert. 1990. "What Is Philology?" *Comparative Literature Studies* 27, no. 1: 21-25. https://www.jstor.org/stable/40246724.

Weatherby, Leif. 2020. "*Critical Response I*. Prolegomena to a Theory of Data: On the Most Recent Confrontation of Data and Literature." *Critical Inquiry* 46, no. 4 (Summer): 891-99. https://doi.org/10.1086/709228.

Ziolkowski, Jan. 1990. "'What Is Philology': Introduction." *Comparative Literature Studies* 27, no. 1: 1-12. https://www.jstor.org/stable/40246721.