Digital approaches to the language of Shakespearean tragedy

Michael Witmore (Folger Shakespeare Library, Washington D.C., USA)
Jonathan Hope (Strathclyde University, Glasgow, UK)
Michael Gleicher (University of Wisconsin-Madison, USA)

This is a pre-print, and pre-proof version of a paper to appear in Michael Neill and David Schalkwyk (eds), *The Oxford Handbook of Shakespearean Tragedy* (Oxford)

If you wish to cite this paper, please do so from the print version.

Interactive versions of Figures 1-4 can be downloaded from http://winedarksea.org/?p=2013

We are living through a revolution in our ability to study Early Modern literature and culture. In January 2015, the Text Creation Partnership (TCP) released around 25,000 texts drawn from Early English Books Online (EEBO) into the public domain. In the past, these texts have been available as page images to subscriber institutions only. Now anyone can download these as fully searchable text files, and with further releases planned, we can look forward to a situation when every person with an internet connection will be able to download and search something close to the entire corpus of surviving Early Modern printed books.[1]

Access to such corpora gives us the chance to consider questions across larger numbers of documents than has been the norm in literary studies. However, investigating at scale necessarily requires different approaches to scholarship: it clearly is not practical to close-read the entirety of a corpus of 25,000 texts. The challenge is twofold: to engage literary specialists whose specialist knowledge of the field is essential to the interpretation of results; and to frame new research questions. To make the most of these new resources, we must integrate traditional literary scholarship with new approaches.[2]

This chapter aims to give an example of such a combined approach. Specifically, we consider the language of a corpus of 554 printed plays from the Early Modern

period.[3] We explore two research questions, each requiring the consideration of a large collection of plays:

(a) is there a distinct 'language of tragedy'?

(b) is there a distinctively Shakespearean language of tragedy?

These are questions that could be considered using traditional literary-critical methods of extensive reading followed by selective quotation and rhetorically-based argument. Our goal is to show how computational and traditional literary techniques can be combined to give better-grounded answers to these questions: computational techniques allow us to compare the language of *all* 554 plays, reliably establishing the groupings of linguistic features which together characterise the language of genres and authors.

As this chapter is focused on showing the potential of our approach, we present our results first, and defer detailed discussion of the methods until later. However an initial outline of our methodology is important, especially for literary scholars intending to make their own use of EEBO-TCP. Briefly, there are four phases to any computational text study such as this:

(1) a corpus of texts is curated: for this study, we assembled a corpus of 554 texts (as described in note 3), removed any language not spoken on stage (such as speech prefixes, stage directions, and so on), and used an automatic spelling moderniser to normalise spelling

(2) some form of measurement of the content of the documents is made: here, we used software to count the frequency of 113 linguistic features in each play

(3) the measurements taken in phase 2 are analysed: in this study we chose to apply standard statistical methodologies which combine the frequency counts of the linguistic features in ways that allow us to 'see' patterns in the occurrence of those features among the documents by plotting them on a two-dimensional chart

(4) we return to individual texts to examine examples of the broader linguistic patterns in situ so that we can attempt to explain their presence using traditional literary reading

Each phase is important, and all must be completed carefully in order to obtain valid results. It is also crucial to remember that we could have made many different choices at phases two and three: choices of what to count, and how to analyse the results. Each choice brings with it a different set of merits and drawbacks, and as more work is done in this new field we will discover more about the best things to count, and the best statistical methods to analyse the results.

Figure 1 shows an initial result from the first three phases of our analysis. Each play is represented by a dot, with tragedies picked-out in black. The dots are positioned such that plays that are linguistically similar to each other are placed next to each other, while plays that are dissimilar are further apart. The statistical method we used to determine similarity and difference is called Principal Components Analysis (PCA), and is designed to summarise as much of the variation in the linguistic measurements we made for each play as is possible in a two-dimensional chart. We have chosen to use PCA in this study as it is a standard approach for summarising such complex data, and allows for visual presentation, but there are many other ways in which we could investigate and visualise our measurements.

Turning to Figure 1, we can see that the tragedies are not evenly distributed amongst the plays: they tend to be more to the left of the graph, and to be more towards the top. Since the process of positioning was 'blind' to the generic labelling of individual plays, the fact that the tragedies differ from the total corpus in this way implies that there are differences between them and the other plays in terms of what *was* considered for the positioning, namely the frequencies of the linguistic features we counted. In brief, Figure 1 suggests that there is indeed a 'language of tragedy': a group of linguistic features whose use

correlates with a play being a tragedy (statistical tests confirm that these differences are not likely to be due to chance).

We can contrast this with Figure 2, where the comedies have been marked in black. Again, the marked plays tend to occupy a specific region of the graph, however comedies occupy a different region to tragedies: they tend to be more to the right, and lower. This result suggests strongly that tragedy and comedy differ in their use of linguistic features – and it invites further exploration. What are the features each genre uses? *Why* do genres use different types of language? Are the plays from each genre on the 'wrong' side of the dividing line experiments, or mixed-genre texts? Do certain writers predictably write 'close to' or 'over' the line? There is not space here to do more than begin to consider some of these questions: they indicate the extent of work for the future.

Figure 1: The 554 play Early Modern Drama corpus with tragedies highlighted in black

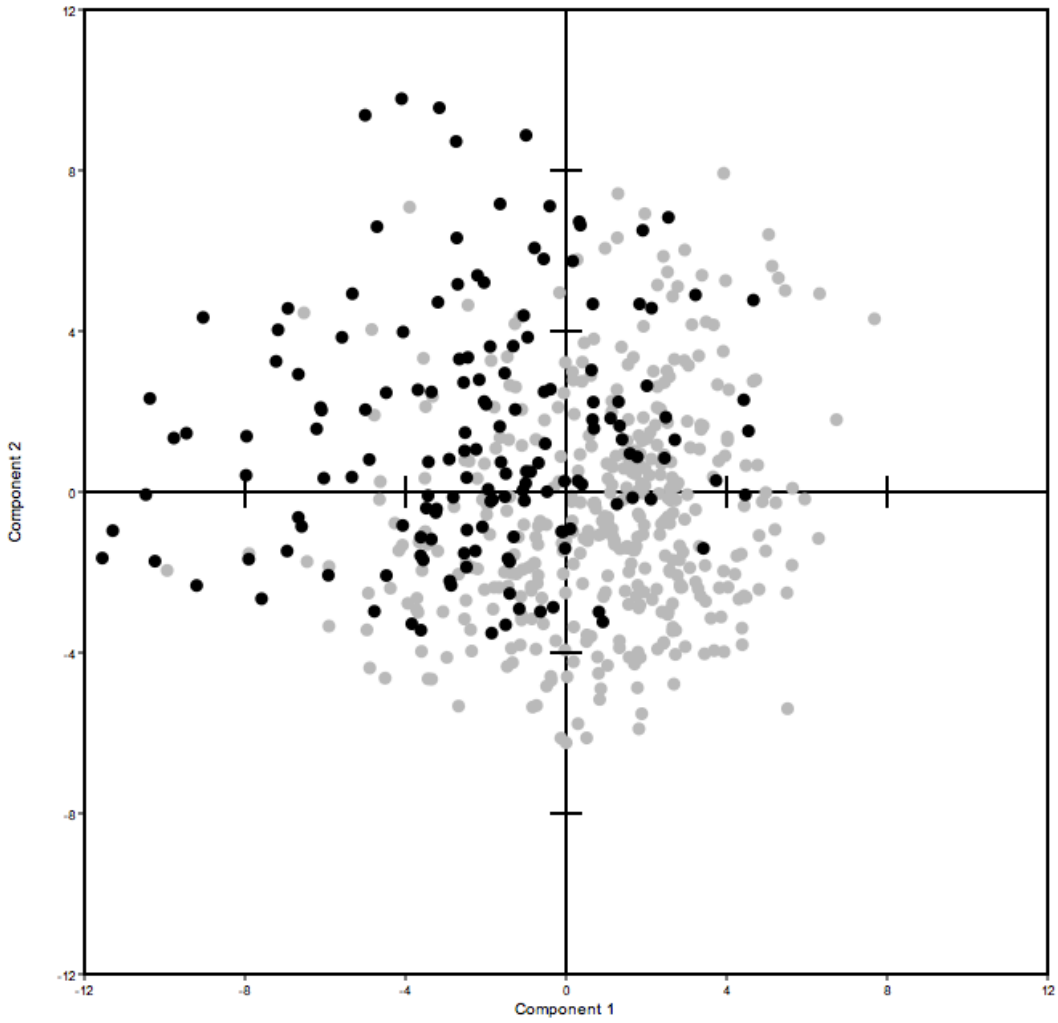
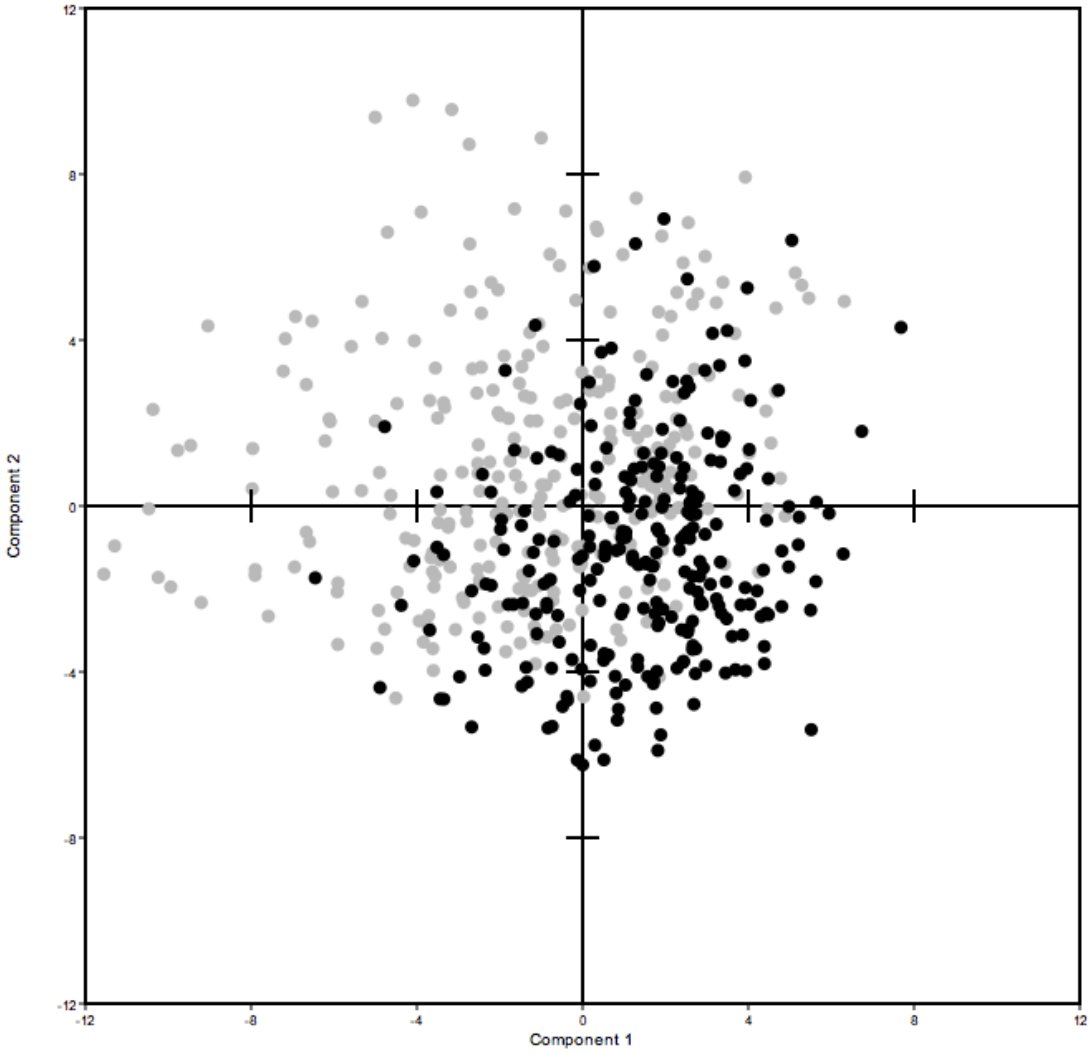


Figure 2: The 554 play Early Modern Drama corpus with comedies highlighted in black

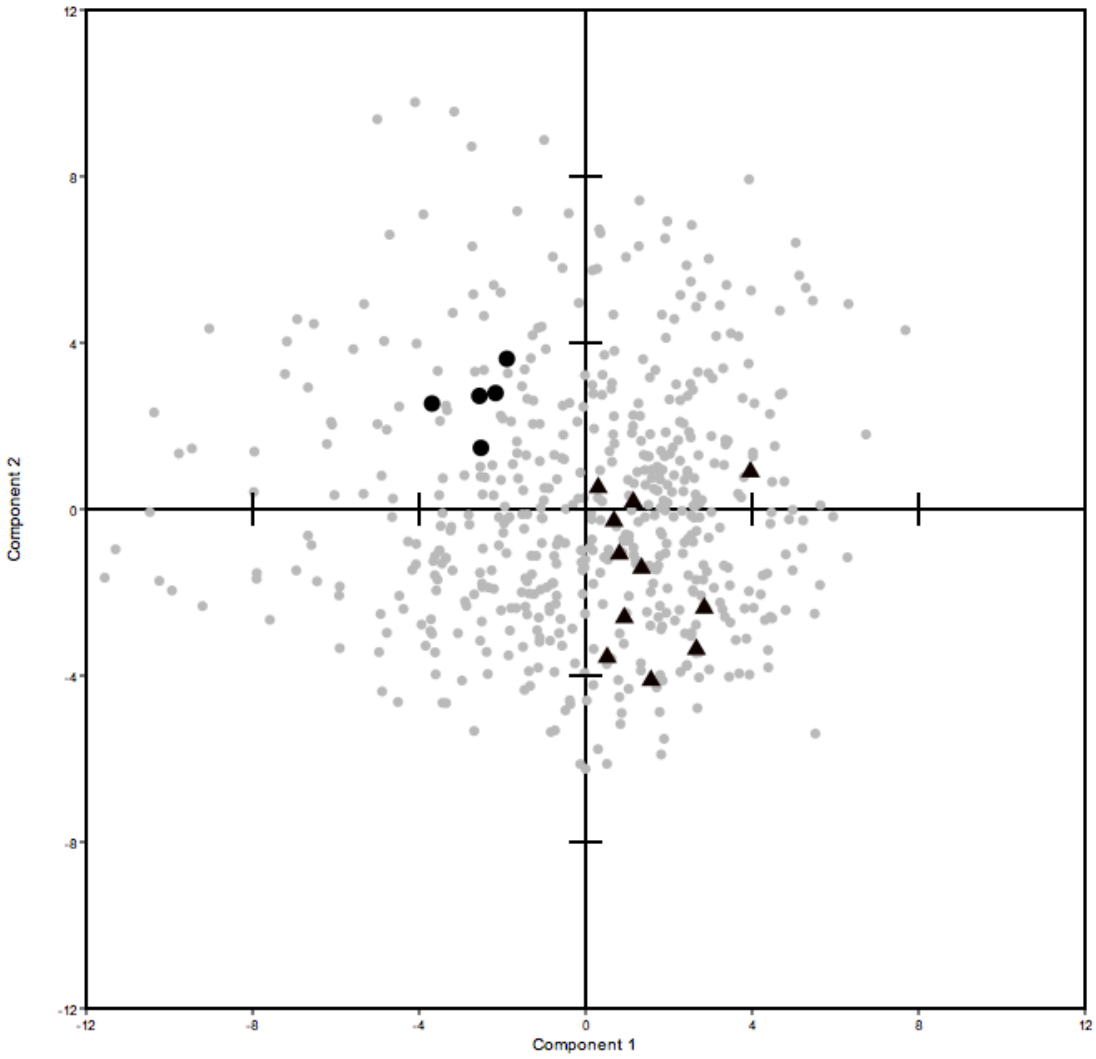


We can also say that there are areas that each genre avoids strongly: tragedies are virtually absent from the bottom right-hand quadrant of the graph, while comedies are very rare in the opposite, top left quadrant. This oppositional patterning tells us that there is an 'anti-signature' for each genre: that is, a set of linguistic features that each genre avoids, but which is present in many of the plays in the other genre. This ability to identify absence as readily as presence is one of the strengths of quantitative-digital work, and one of the things it does that traditional human readers are less good at.

So what are the characteristic languages of tragedy and comedy in the Early Modern drama corpus? What differences in linguistic practice are reflected in these separations in the space of the graph? On analysis, it turns out that tragedies (perhaps not surprisingly) favour a set of linguistic features used to communicate negative emotion and affect, while comedies, less predictably, favour a set associated with the representation of rapid, highly interactive speech, including first and second person pronouns, questions, discourse markers, words for social roles and relationships, and imperatives. We term these groups the 'negative' and 'oral' groups.

To show these linguistic styles in context, we will use the work of George Chapman, who emerges from this study as a generically exemplary writer.[4] Chapman's plays are visualised in Figure 3, with his tragedies picked out as black circles, and his comedies as black triangles. This graph repeats the pattern of generic separation between tragedy and comedy observed in Figures 1 and 2: Chapman's tragedies all group to the upper left, in the quadrant which has mainly tragedies, and very few comedies (and which we can term the 'core' tragedy space). Chapman's comedies, although slightly less tightly grouped than his tragedies, all lie to the right and lower, and almost all are in the quadrant of the graph which has many comedies and almost no tragedies (the 'core' comedy space).

Figure 3: George Chapman's plays in the Early Modern drama corpus (tragedies = black circles; comedies = black triangles)



Before we give textual illustrations of these contrasting strategies, a caveat. Traditionally, literary criticism has worked by building arguments around exemplary quotations, chosen for their rhetorical force. The quotations represent the key moment, or the height of the author's performance. The argument stands or falls on the aptness and quality of the supporting quotations. Russ McDonald's essay on the language of tragedy is an outstandingly good example of this (in both senses: it is a good example, and it is an excellent essay).[5] McDonald uses skilfully chosen quotations to make his points, working from the particular to the exceptional. What we are doing here is something very different. We are counting at scale (554 plays), and we are not looking for unusual, stand-out passages, but for large-scale patterns that assert themselves repeatedly over many plays. We are comparing large groups of plays on the basis of what is similar between them – and the differences we identify are similarly at scale. Things that happen only once, or only a few times, do not register in our analysis: they are drowned out by the force of numbers. When we present you with quotations from plays, they are intended to be representative of what goes on across very large amounts of text: they are not 'plums' plucked from a literary pie, but are the stodge that makes up the vast majority of what is going on in the text. This is an unfamiliar way of thinking about literary practice, and one of the challenges for literary studies in the coming years is to come to terms with it as a method.

Here is a passage from Chapman's tragedy *Bussy D'Ambois* with negative group features in bold:[6]

will she but disclose
Who was the **hateful** minister of her love,
And through what maze he served it, we are friends.
It is **a damned** work to pursue those secrets,
That would op more **sin**, and prove springs of **slaughter**;
Nor is it a path for Christian feet to touch;
But out of all way to the health of souls,
A **sin impossible to be** forgiven:

Which he that dares commit;
Good father cease:
Tempt not a man **distracted**; I am apt
To **outrages** that I shall ever rue:
I will not pass the verge that bounds a Christian,
Nor **break the** limits of a man nor husband.
Then God inspire ye both with thoughts and deeds
Worthy his high respect, and your own souls.
Who shall remove the mountain from my heart,
Op the seuentimes-heat furnace of my thoughts,
And set fit outcries for a soul **in hell**?
O now it nothing fits my cares to speak,
But thunder, or to take into my throat
The trump of Heaven; with whose determinate **blasts**
The winds shall burst, and the **enraged** seas
Be drunk up in his sounds; that my hot woes
(**Vented** enough) I might convert to vapour,
Ascending from my **infamy** unseen;

George Chapman, *Bussy D'Ambois* TCP A18403

And here, by contrast, is a passage from the comedy *An Humorous Day's Mirth* with oral group features in bold:[7]

Honour to my good lord, and his fair **young lady**.
Now **Monsieur** Satan, **you are** come to tempt and prove at full the spirit of **my wife**.
I am **my lord**, but vainly **I** suppose.
You see she dares put on this brave attire fit with the fashion, which **you think** serves much to lead a **woman** into light desires.
My lord I see it: and the sight thereof doth half dismay **me** to make further proof.

Nay prove her, prove her **sir**, and spare not**: what** doth the witty **minion** of our **King** think any dame in France will say him nay**?** but prove her, prove her, see and spare not.
Well **sir**, though half discouraged in **my** comming, yet Isle go forward: **lady**, by your leave.
Now **sir**, **your** cunning in a Ladyesproofe.
Madam, in proving **you** I find no proof against **your** piercing glancings, but swear I am shot thorough **with your** love.
I do believe **you**: who will swear he loves, to get the thing he loves not**?** if he love, what needs more perfect trial**?**
Most true rare **lady**.

George Chapman, *An Humorous Day's Mirth* TCP A18419

Clearly these are very different discourse situations: the conventions of the genres, and the demands of the different narrative structures they set up, encourage the use, and avoidance, of the associated features. Tragedies favour relatively formal speech situations, typically with longer speeches, fewer exchanges, and with declarative, rather than interactive, tendencies. Comedies favour informal situations, with rapid shifts of speaker, and the use of features to mark turn-taking, attention-getting, contradiction, agreement, and so-on. Now, a reader not well-disposed to quantitative work might observe that it is not surprising to find that tragedies favour language about death, sorrow, and nasty things in general, and that we hardly need computers and advanced statistical analysis to point this out to ourselves – and we would have to agree. But there is something, if only relief, in new techniques that produce results that make sense to domain-specialists. This, at the very least, can give us some confidence that the linguistic software is counting things that do have a role in producing recognisable literary effects.

We can also say that quantitative methods add several things to our understanding we could not otherwise arrive at. For one thing, we now know that these differences are visible across the whole range of Early Modern drama:

the human cultural categories of tragedy and comedy are mapped by consistent differences in the use of a range of quantifiable, small-scale linguistic features. We have linked high-level, rather abstract conceptions (genre categories) with low-level linguistic forms. The statistics link these in a correlation (high use of negative features correlates with plays which have been termed tragedies), and human researchers can offer plausible explanations for why there might be an association.

It is important to note that simply establishing a statistical correlation between a group of texts and a set of linguistic features does not guarantee that we have found something that is significant in a literary sense. Statistical significance does not equate to literary significance or interest. The statistics simply form the basis for literary investigation and interpretation: they are a starting point, and a way to return to the texts with a new perspective, but there are many correlations in our data that we do not find 'interesting' in a literary sense.

Furthermore, we can note that quantitative methods allow us to be specific about what tragedies are *not* in a way traditional literary analysis would find much harder. This is important, because figure 1 shows tragedies spread across a large area of the graph, implying that there is a large amount of linguistic variation within the genre. Tragedies plotted in the upper right of the graph will have a different linguistic make-up to those in the lower left – and tragedies close to the centre of the graph will differ from those in the outlier cloud. So what characterises tragedies is not simply the shared *presence* of a set of features (the negative ones listed above), but also the shared relative *absence* of a second group (the oral features).

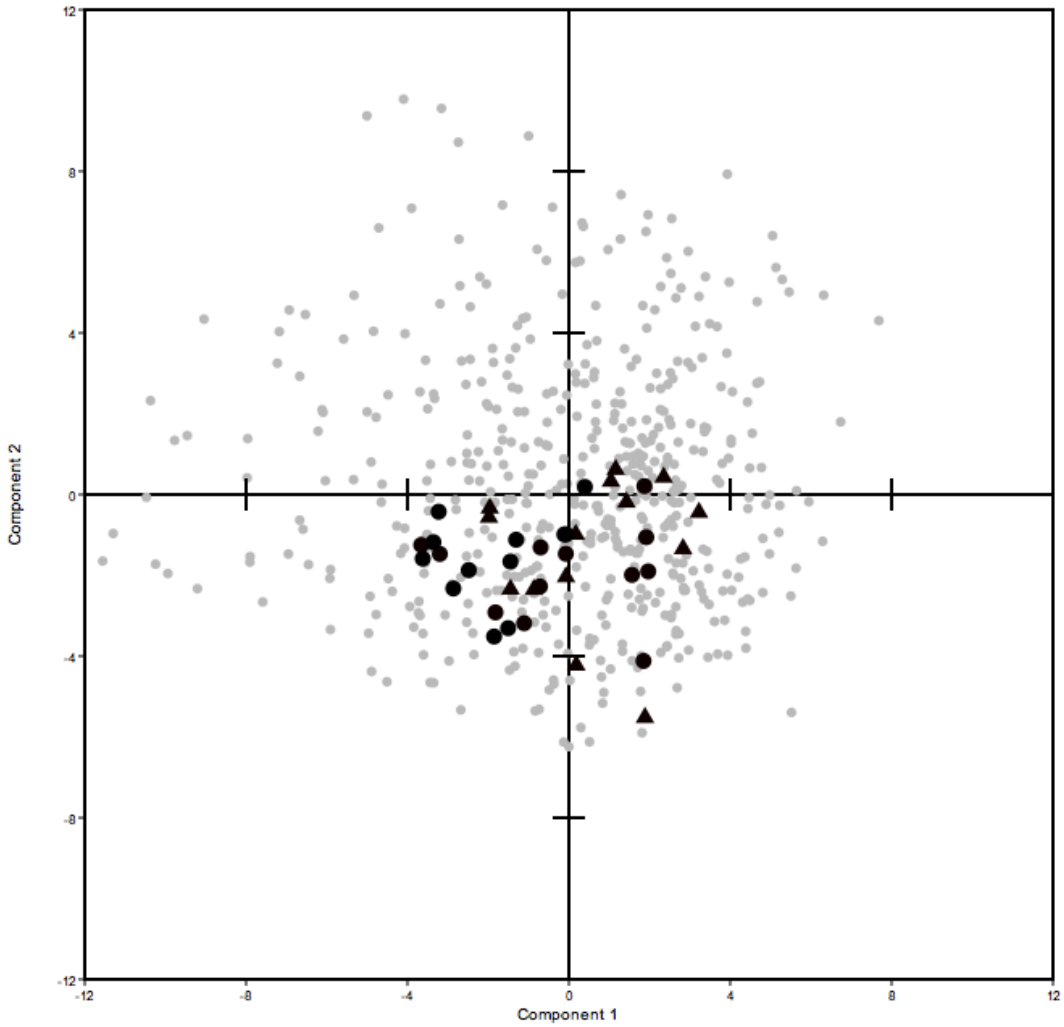
A further strength of quantitative methods is that they can often present us with results that are counter-intuitive and which open up avenues for further research. In the course of this study, for example, we identified a third set of associated features with a role in patterning the placement of the texts, whose distribution raises some fascinating literary questions. Features in this group share a function in representing the self.[8]

Counter-intuitively, to us at least, this 'self' group is generally more characteristic of comedies than tragedies: use of this group of features tends to pull plays into the lower right-hand side of the graph. We find this counter-intuitive because of literary accounts of tragedy as the locus for Renaissance investigation, even creation, of the self: if these were correct, we might expect 'self' features to be a characteristic of tragedy, or at least of Shakespearean tragedy. As we will see, however, this is not the case. We do not have space in this overview of the corpus to investigate this finding fully, but we will discuss it further in relation to Shakespeare below, and we invite other scholars to test our finding.[9]

To sum-up the first part of our study. The 'negative'/'oral' opposition is the key distinction between the language found in tragedies and that found in comedies. Of course, both genres use features from both groups: tragedies use first person pronouns, and comedies use negative language. But what digital analysis can show us are the consistent differences in rates of use: overall, tragedies consistently use more negative language, and less oral language, than comedies. Look, like a human, at any one speech, and features from both groups are likely to be present; look, like a computer, at 554 plays at once, and small but consistent differences in frequency combine to produce clear tendencies.

We now move on to consider Shakespeare. Figure 4 once again shows the entire Early Modern drama corpus, this time with Shakespeare's tragedies picked out as black circles, and his comedies as black triangles. Figure 4 makes an interesting comparison with Figure 3: where Chapman's tragedies and comedies divided neatly into the tragedy and comedy quadrants, Shakespeare's plays are not so well-behaved.

Figure 4: Shakespeare's plays in the Early Modern drama corpus (tragedies = black circles; comedies = black triangles)



The most striking difference between Figures 3 and 4 is the lack of clear generic separation between Shakespeare's tragedies and comedies. Where Chapman's tragedies and comedies grouped by genre in different parts of the graph, Shakespeare's plays all occupy broadly the same space. Additionally, both genres tend to lie slightly away from their 'core' quadrants as identified in the corpus-wide analysis. This is particularly clear for tragedy, pulled lower and to the left of the core tragedy space, and slightly less clear for comedy. This is a suggestive finding: on the one hand, Shakespeare's plays are all within the central space of Early Modern drama, so his linguistic choices are not extreme, but on the other, his linguistic choices map slightly to the margins of the core, with less clear generic distinction than we find in some writers (e.g. Chapman). This suggests a mixing of features in general, though it is important not to forget that Shakespeare's plays do still show usage of the typical genre-features at rates comparable to those of other 'core' writers (if this were not the case, his plays would lie in the outer regions of the graph).[10]

If we look at Shakespeare's tragedies, and ask what sends them to the lower left, we can say that this is primarily due to the presence of three groups of language-features. First, Shakespeare's tragedies have the expected higher rates of the 'negative' group features. Here, as an example, is a passage of 'negative' language from *Titus*:

Art thou not **sorry for** these **heinous** deeds?
I, that I had not done a thousand more:
Even now I **curse the** day, and yet I think
Few come within few compass of my **curse**,
Wherein I did not some **Notorious ill**,
As **kill** a man, or else devise his **death**,
Ravish a Maid, or plot the way to do it,
Accuse some Innocent, and **forswear** my self,
Set **deadly Enmity** between two Friends,
Make **poor** men's Cattle **break their necks**,
Set **fire on** Barnes and Haystackes in the night,

And bid the Owners quench them with the **tears**:
Oft have I dug up **dead men** from their graves,
And set them upright at their dear Friends door,
Even when their **sorrows** almost was **forgot**,
And on their skins, as on the Bark of Trees,
Have with my knife carved in Romaine Letters,
Let not your **sorrow** die, though I am **dead**.

William Shakespeare, *Titus Andronicus* TCP A11954_27

In addition to this, however, Shakespeare's tragedies are pulled down into the lower part of the graph because they use two sets of features more than those tragedies plotted in the mid-point, and upper right of the graph. The first set is the 'oral' group (primarily associated with comedies. Here is the opening of *Julius Caesar,* which has a large amount of 'oral' language:

HEnce: home **you** idle Creatures, **get you** home:
Is this a Holiday**?** What, know **you** not
(Being Mechanical) you ought not walk
Upon a labouring day, without the sign
Of **your** Profession**?** Speak, what Trade art **thou?**
Why Sir, a Carpenter.
Where is thy Leather Apron, and **thy** Rule**?**
What dost **thou** with **thy** best Apparel on**?**
You sir, what Trade are **you?**
Truly Sir, in respect of a fine Workman, **I am** but as you would say, a Cobbler.
But what Trade art **thou?** Answer **me** directly.
A Trade Sir, that **I hope I** may use, with a safe Conscience, which is indeed Sir, a Mender of bad souls.
What Trade **thou** knave**? Thou** naughty knave, what Trade**?**
Nay I beseech **you** Sir, be not out with me: yet if **you be** out Sir, I can mend **you**.

What mean **thou** by that**?** Mend **me**, **thou** saucy Fellow**?**
Why sir, Cobble **you**.
Thou art a Cobbler, art **thou?**

William Shakespeare, *Julius Caesar* TCP A11954_30

Admittedly this is an extreme example, but the presence of all of Shakespeare's tragedies below the horizontal axis of the graph shows that this increased use of 'oral' features (relative to tragedies plotted above and to the right) is consistent.

The second set of features pulling Shakespeare's tragedies away from the core tragedy space is an additional group termed the 'world-space' group. The linguistic items in this group are concerned with describing things in the visible world, and mapping the spaces in which objects and people exist.[11] One implication of the location of Shakespeare's tragedies in the lower left of the graph is that they are more concerned with representing the 'real' physical world than many other Early Modern tragedies (especially those in the upper right of the graph).

Here is an example of the 'world-space' group from *King Lear*:

Thou were better **in a Grave**, then to answer with thy uncovered **body**, this **extremity of the Skies**. Is man no more then this? Consider him well. Thou ow'st the **Worm** no **Silk**; the Beast, no Hide; **the Sheep**, no **Wool**; **the Cat**, no **perfume**. Ha? Here's three on's are sophisticated. Thou art the thing it self; unaccommodated man is no more but such a poor, **bare**, **forked** Animal as thou art. Off, off you Landings: Come, unbutton **here**.
Prithee Nunckle be contented, it is a naughty night to **swim** in. Now a **little fire** in a wild **Field**, were like an **old** Lechers **heart**, a **small spark**, all the rest on's **body**, **cold**: Look, here comes a **walking fire**.

William Shakespeare, *King Lear* TCP A11954_33

There are two surprising findings here. First, Shakespeare's tragic language has an increased focus on the external, physical world compared to the language of most other Early Modern tragedies. Second, Shakespeare's tragedies show an *avoidance* of the 'self' group. High use of the 'self' group would have pulled Shakespeare's tragedies up and to the right, but they are located diagonally opposite, low and to the left – so within tragedies as a whole, Shakespeare's plays use these features less than the average.

Figure 4 shows that while there is generally not a very clear generic separation at this scale between Shakespeare's comedies and tragedies, there is a clear tendency for only comedies to have positive values on the horizontal axis: out of the Shakespeare corpus, only comedies appear on the right hand side of the graph. This means that Shakespeare uses 'self' language relatively more in comedy than in tragedy (though there are some comedies which do not use it very much). Here is an example of 'self' language from *The Winter's Tale*, as Autolycus slyly misrepresents himself.

Doest lack any money? **I have** a little money for thee.
No, good sweet sir: no, I beseech you sir: **I have** a Kinsman not past three quarters of a mile hence, unto whom **I was** going: **I shall** there have money, or any thing **I want**: Offer me no money I pray you, that kills **my heart**.
What manner of Fellow was hee that robbed you?
A fellow (sir) that **I have known** to go about with Troll-my-dames: **I knew** him once a servant of the Prince

William Shakespeare, *The Winter's Tale* TCP A11954_14

And here is an example of 'self' language in the Shakespeare tragedy that lies closest to the right hand side of the graph: *Julius Caesar.* This comes from Brutus' reply to the conspirators, employing 'Self Disclosure' ('I am', 'I have', 'I would') and 'Metadiscourse' ('I shall', 'I will consider'):

That you do love me, **I am** nothing jealous:
What you would work me too, **I have** some aim:
How **I have** thought of this, and of these times
I shall recount hereafter. For this present,
I would not so (with love I might entreat you)
Be any further moved: What you have said,
I will consider:

William Shakespeare, *Julius Caesar* TCP A11954_30

There is of course much more to explore in these relationships. What we have provided here is the beginning of an outline of a space mapped out using one set of criteria.[12] This space can be investigated much more thoroughly – and other spaces can be created by counting different features in the same set of plays. In the following section we provide more detail of our methodology, a discussion of some of the implications for literary study, and a summary of findings and further questions.

Method

We use Docuscope, a rhetorical analysis package, to count linguistic features in a corpus of Early Modern drama. Computationally, Docuscope is a very simple string-matching program. It searches text files for strings of characters (made up of single words, phrases, and in some cases punctuation marks) which it recognises from its dictionaries (which are simply lists of words and phrases). Each word or phrase that is recognised is tagged as belonging to a 'Language Action Type', or LAT. This is where Docuscope becomes more sophisticated: the LATs are rhetorical-linguistic categories, constructed and populated by the human designers of Docuscope. The LATs attempt to capture words and phrases that have predictable effects on the reader of a text: that create different experiences when reading.[13] We could have characterized the differences between genre with less developed categories – counting, for example, all of the prepositions in the plays and using these as discriminating features. We find, however, that the Docuscope categories are more 'interpretable' because they are functionally driven (and human-crafted).

For example, the LAT 'First Person' tags first person pronouns ('I', 'me' and so on): a text high in 'First Person' is likely to be presenting a relatively straightforward set of self-references. This may seem a rather obvious and crude example, but some of the subtlety of Docuscope's categories (and one of the drawbacks in its method of operation) can be seen by comparing 'First Person' with three other LATs, 'Self Disclosure', 'Self-Reluctance', and 'Autobiography'. These three LATs attempt to tag more complex forms of self-representation. 'Self-Disclosure' tags first person pronouns in combination with verbs associated with self-revelation, or prepositions doing a similar job ('I am', 'I think', ''I feel', 'I believe', 'I confess', 'to me', 'for me') – here, a more complex, more conscious, self-representation is effected. 'Self-Reluctance' tags first person pronouns in combination with verbs of resistance or disagreement (''I regret that', 'I am forced to', 'I had to', 'against my will') – again attempting to capture a more nuanced self-representation. 'Autobiography' tags first person pronouns combining with verbs, nouns, or conjunctions in self-representations which

reflect back in time on a personal past ('I have been', 'I was', 'when I', 'my name', 'my daughter'). This gives some sense of the distinctions the designers of Docuscope were seeking to be able to make in textual effects. It also demonstrates one of the restrictions on Docuscope's counting: no character string can be tagged more than once, and each character string is included in the longest string Docuscope can find. So any first person pronoun tagged as 'Autobiography' is not available to be tagged as 'First Person'. The same is true for all character strings: Docuscope behaves as if all elements of language contribute once only, and in only one way, to the effect of a passage. This makes counting simpler, but it certainly runs against what we know of language.

The version of Docuscope used in this study has 113 categories into which it places strings: for each text it counts, it produces a tagged version of the text (viewable in a text-viewer), and a spreadsheet-readable file (as 'csv' or comma-separated variable file). The csv file gives the normalised frequency of each of the 113 LATs for each of the texts in the analysed corpus (in our case 554 plays).[14] We discard any LATs where the frequency is zero, or close to zero. In this study, our results are based on a spreadsheet with 554 rows and 72 active columns.[15] Once we have passed our play corpus through Docuscope, we have effectively compared 554 plays on the basis of 72 points of comparison – giving us 39,888 data points. At this point, we come up against the limitations of human attention and cognition. We could try and 'read' the 39,888 cells of the spreadsheet, looking for patterns of similarity and difference, perhaps including extra columns of metadata (author, date, genre, and so on). Of course, we would not get very far: humans are bad at reading large tables of numbers, and in any case, a lot of the information in the spreadsheet is not very interesting: all the plays may be more or less the same on a particular LAT, or the variations in frequency between them might have no pattern.

So what we need to do is reduce the complexity of the information, and present it in a way that will enable us to see patterns of similarity and difference that are interesting to us. This is the basis of almost all statistical analysis and

visualisation: *reduction* in information, and then presentation in a form humans can cope with cognitively.

In fact, put like this, what we are doing is pretty similar to what literary critics have tended to do in the past: they reduce the complexity of the information by focussing on a few attributes of the texts they study, and they make a huge body of material cognitively accessible to humans by citing just a few quotations in evidence. And the ultimate aim can be cast in similar terms too: both approaches seek to say 'look at this text – it is similar to this text for these reasons, and different to those texts for these reasons'. This is a simplification of several hundred years of literary criticism of course, but it is not a gross misrepresentation.

How much of a simplification we make to our data during the analysis will become clear as we look more closely at what we have done by counting 72 LATs in each of our plays. What we effectively did was to plot all 554 plays in a space defined by the total linguistic variation between all the LATs in the corpus. Each play is located at a unique point in that space – and each point is fixed by a set of co-ordinates made up by all of the frequencies for each LAT in the play. Because we counted 72 LATs, each play has 72 co-ordinates: and the space representing the linguistic variation in the Early Modern drama corpus is made up of 72 dimensions.

Once again, we are beyond the limits of human brains. We can easily imagine a one dimensional space: that is a line. We can plot all 554 plays along a line using the frequency results from any one of the LATs, running from lowest to highest. This would be a useful visualisation, as it would quickly show us the relationships between plays for this single LAT. We can also easily imagine a two dimensional space: this is a graph with each axis representing the results from a different LAT. Each play would have a position in the space of the graph produced by the two values. Again, this is a useful visualisation – perhaps even more useful – as it shows us the relationship between two LATs: plays high on both will be at the top right of the graph; plays low on both at bottom left. And

we can also imagine a three-dimensional visualisation, adding a third axis, and LAT, to our graph, so the plays are arranged in a cuboid space.

At this point, while our brains give up, mathematics does not. We can carry on adding LATs and axes until we have 72 dimensions, with our plays arranged within the resulting, unimaginable, but mathematically describable, space. What we now need is a means of seeing the 72-dimensional space in a form humans can process: a way of reducing the dimensionality. There are various standard ways of doing this statistically, and in this study we have used Principal Components Analysis.[16] This is a common technique for reducing the dimensionality of complex data sets, and revealing patterns of association within high-dimensional space. Like all statistical techniques, it is simply one way of slicing through the data: it does not give the only possible view of a data set, and there may be other approaches that will show other things about the data. We have used PCA here because we think the results are interesting – and we have attempted to support them from non-PCA techniques[17] – but once again, readers should treat this study and these results as experimental and exploratory. There is much more to be investigated in this data set, and many other approaches to be tried.

The graph shown in figures 1-4 is a two-dimensional representation of the 72-dimensional data space. The axes of the graph are two 'principal components', or PCs. These are derived from a mathematical reduction of the relationships between the 72 LATs in Principal Components Analysis space (PCA Space). Each PC is an attempt to represent as much of the variation in the 72-dimensional data space as possible: the 72 coordinates fixing each play in PCA space are reduced to two, one on each PC, which reduce and summarise the much more complex high-dimensional space. If we imagine the 72-dimensional space as a graph with 72 axes, all pointing in different directions, then the PCs are attempts to draw axes which 'summarise' as many as possible of the 72 axes, re-orienting the data around a smaller number of axes that capture the most variation. Together, the two PCs we have extracted from this data set account for just over 26% of the total variation in the drama corpus. This means that we have thrown out about

74% of the information! The reward for doing this, we hope, is that we can now see relationships and patterns in the data that are impossible to visualise at 72 dimensions: but we need to remember that this is a *reduction* – a simplification – and that other views and representations of the data set are possible.

Concluding discussion:

Digital tools and resources (such as EEBO-TCP) allow literary scholars to approach their object of study in new ways. In many respects, we continue to do the same thing: we read texts and compare them, making links and contrasts. We associate certain texts, and we separate others. Traditionally, literary criticism has made these associations and separations using notions such as 'genre', 'influence', and 'period'. The evidence for links and distinctions has been gleaned by close reading, and is constituted by quotation, and summary of plot, theme and technique. Although not always the case, it is generally true that literary criticism has sought out the exceptional above the typical: a tragedy is discussed as such because of an explicit, or implicit, claim that it is an outstandingly good tragedy, or is unusual in some important way. Plays considered to be 'average' or 'typical' examples of their genre may be referenced as such, but are unlikely to be the focus of sustained examination.

As Ted Underwood has suggested, the dominant model in literary study is one of exceptionalism and fracture: the narratives literary scholars construct stress the turning point, the break with the past: major writers are points of sudden change, after which nothing is the same.[18] While digital methods allow us to continue comparing and making claims about similarity and difference, the fact that they allow us to do these operations at scales, and with a consistency not possible for single human readers, shifts the nature of literary study as individual texts are read through the lens of much larger groups. Further, as Underwood has argued, the nature of the evidence digital studies present for use in constructing literary arguments, and the stories that evidence tends to tell, re-orients the history of literary study itself: digital evidence, high in frequency, stressing the average and the mean, rather than the exceptional outliers, running over timescales and including volumes of text impossible for single human readers, tends to emphasise gradual change and continuity rather than sudden fracture. The differences it detects tend to be relative rather than absolute.

Digital methods can detect things human readers find impossible or very difficult: shifts in the frequency of very common features[19]; absences rather than presences. And these offer new ways of approaching familiar texts – new ways of contextualising them. Digital methods also offer an opening up of the canon: with large corpora, each text is treated equally by the software – associations between canonical texts and those less, or hardly ever studied, are possible. Texts previously available only in research libraries, or subscription-only web resources, and hardly mentioned in accounts of literary history, or on undergraduate survey courses, will pop up next to canonical texts in visualisations: students will be able to access them, and will have a reason to do so.

However, we should note that there are downsides to this apparently bright prospect. In digital work we are constrained by other things than the limits of human attention spans and cognition – though it is easy to forget those constraints as we conjure multi-coloured three-dimensional graphs from our software. We are constrained first of all by what *can* be counted: at a base level this is strings of characters, or, slightly more sophisticated, tags in a text file. When we claim to be counting rhetorical features, or influence, or style, we are really counting something we have identified as a proxy for those things: something that can, ultimately, be reduced to a set of character codes a program can recognise. Our results will only be as good as the relationship of the proxy to whatever it is we think we are studying: identifying and describing the proxy is a job for literary specialists, as Underwood has argued – and often the process of identifying such proxies prompts fundamental questions about the object of study: what is 'influence'? what is 'style'? These questions are outside the remit of digital methods; they belong squarely back in 'traditional' literary theory.[20]

Even assuming some ideal situation where we identify a perfect, countable, proxy for whatever it is we wish to study, we are also constrained by what we have to count *in*. The impressive size of digital data sets offers an illusion of completeness: EEBO-TCP offers us 'every' printed text in English between certain dates. But of course, this is 'really' the set of *surviving* printed texts from

the period[21] – and, while an impressively extensive collection, it is not *complete* even of those printed texts we know to have survived – 'new' texts are still being added to EEBO. The huge body of surviving Early Modern manuscript material is absent – which is significant for this chapter given the number of manuscript plays we know about in the period (and even more so for those who wish to use EEBO-TCP for cultural and linguistic history).

Even within the texts we have, we must remember that no data set is perfect: EEBO-TCP texts are human artefacts: keyed transcriptions of microfilm; microfilms which are themselves imperfect representations of imperfectly printed texts. The EEBO-TCP text files have many gaps, marked by the transcribers, where they simply could not read what was in front of them (this is especially the case for black letter texts). We can hope that these gaps are (more or less) evenly distributed through the corpus, and that the counts we perform are on features so frequent that the losses due to miss- or missing transcription are negligible – that the data is 'good enough', in a telling statistical phrase likely to strike literary, and especially textual scholars, as chilling – but we must always be prepared to accept that our results may be affected by such gaps.

Summary of findings and further questions

1 It is possible to identify linguistic signatures for tragedy and comedy: compared to the corpus of Early Modern drama as a whole, tragedies are characterised by increased use of language involved with the communication of negativity, and reduced use of language involved in representing oral exchange. (Comedies show the reverse pattern.)
Further questions: (a) Why do writers follow these linguistic patterns when writing in these two genres? Is there something about the dramatic situations the genres create that favour the use of certain types of language over others? (b) Which plays or writers go against these general trends to produce tragedies which are mapped onto the 'comic' half of the graph?

2 Shakespeare's tragic language, while falling in the central 'core' of linguistic practice, can be characterised in relation to the tragedy corpus as a whole as using more language associated with real-world description, and the representation of oral exchange. Perhaps surprisingly, given the history of critical comment on Shakespeare, it does not show an increase in the use of language associated with self-revelation.
Further questions: (a) Why does Shakespeare move towards real-world reference and orality? (b) Is our surprising finding regarding 'self' language robust? Are we counting the right features? What constitutes 'self' in language and can it be counted? (c) If our finding that Shakespeare fits within the common core of writers in terms of linguistic practice is right, what makes him 'better' than other writers who share his linguistic practices?

3 Taken as a whole, the corpus of Early Modern drama is characterised by a central core group of plays with a broadly similar linguistic makeup.
Further questions: (a) Is this core group made up of particular types of plays, authors or genres – for example, are professional playwrights/major companies located here?

4 Outside this core group, outlier plays are found in only certain parts of the possible linguistic space (on the left-hand side of a diagonal line drawn across the graph).
Further questions: (a) Is the outlier cloud made up of particular types of plays, authors or genres? (b) Why do no plays have the linguistic makeup that would put them on the right-hand side of the diagonal?

5 There are many practical issues to be faced as we develop this hybrid approach to exploring literary texts. Notably in this paper we have come up against the problem of referencing within the processed TCP texts we have used for our analysis. Scholars will need to address this if we are to be able to move through the TCP text set easily: especially as we move from working on the relatively well-known drama texts to the rest of the corpus.

Notes

1 For the TCP project, see http://www.textcreationpartnership.org/tcp-eebo/ (accessed 4.5.2015). EEBO-TCP Phase I contains 25,363 texts, manually keyed to allow full text searching. EEBO-TCP Phase II aims to similarly transcribe a further 45,000 texts from EEBO. Files can be downloaded from https://github.com/textcreationpartnership (accessed 4.5.2015), and a master list of all TCP files is at https://github.com/textcreationpartnership/Texts/blob/master/TCP.csv (accessed 4.5.2015). EEBO-TCP seeks to cover as much as possible of the surviving corpus of Early Modern Print, but we should remember that a large number of texts have been lost (see note 21) – and the process of transcription is continuing, with new files regularly added to the corpus.

2 The authors of this paper are involved with the Mellon-funded Visualising English Print project, which is developing tools and methods for analysing EEBO-TCP (and any large corpus of texts). See the project website http://graphics.cs.wisc.edu/VEPsite/ (accessed 4.5.2015) and project members' blog http://winedarksea.org (accessed 4.5.2015)..

3 Our corpus of dramatic texts comes from the EEBO-TCP transcriptions, and was originally selected and supplied to us by Martin Mueller, for which we are very grateful. Subsequently, in order to ensure that the entire corpus was processed in the same way, we re-selected and re-processed EEBO-TCP texts. Texts went through three stages of processing: (i) we performed automatic clean-up to remove certain characters introduced during transcription (for example, the pipe character < | > frequently appears where hyphens have triggered an incorrect word division in the transcription); (ii) we modernized texts automatically, using VARD (http://ucrel.lancs.ac.uk/vard/about/ - accessed 8.5.2015); (iii) we stripped texts of all non-spoken elements (stage directions, act and scene numbers, speaker designations) using XML codes. Genre labels were assigned by Jonathan Hope, drawing on metadata supplied with some of the texts, titles and title pages, and labels given in the Database of

Early English Playbooks (DEEP - http://deep.sas.upenn.edu - accessed 4.5.2015). The metadata originally associated with the play texts, and included in the file TCP.csv (downloadable from the address given in note 1) has been checked, corrected and extensively expanded by Beth Ralston as part of the VEP project. This new metadata, a list of all the plays in the corpus, their genre labels, and the frequency scores used in this study, can be found in a new spreadhseet which we are making available along with the stripped, VARDed .txt files and tagged HTML files of the corpus play texts. For details of this material, see http://winedarksea.org/?p=2013 (accessed 12.5.2015). See note 6 for further comment on the texts and our method of referencing.

4 The plays in our Chapman corpus are as follows (eleven comedies first, followed by five tragedies). Note that dates and ascriptions are from our corrected metadata referenced in note 4 above, but many dates and ascriptions in the corpus are conjectural and subject to change in the light of future scholarship.

TCP number	title
A09134	*Fedele and Fortunia* (1585)
A18402	*The Blind Beggar of Alexandria* (1596)
A18419	*An Humorous Day's Mirth* (1597)
A18400	*All Fools* (1601)
A01911	*Sir Giles Goosecap* (1602)
A18413	*The Gentleman Usher* (1602)
A18415	*May Day* (1602)
A18426	*The Widow's Tears* (1604)
A69093	*Monsieur D'Olive* (1605)
A18407	*Eastward Ho* (1605)
A18423	*Two Wise Men and All the Rest Fools* (1619)
A18403	*Bussy D'Ambois* (1604)
A18425	*Caesar and Pompey (Wars of Caesar and Pompey)* (1605)
A18404	*The Conspiracy of Charles Duke of Byron* (1608)
A18404	*The Tragedy of Charles Duke of Byron* (1608)
A18421	*The Revenge of Bussy D'Ambois* (1610)

5 Russ McDonald, 2006, 'The language of tragedy', in Claire McEachern (ed.), *The Cambridge Companion to Shakespearean Tragedy* (2nd ed., CUP), pp. 23-49.

6 We quote directly from the EEBO-TCP texts as processed according to the procedure outlined in note 3. This means that there are no speech prefixes or stage directions. For each text we give a TCP number, which identifies the TCP text file containing the text (if, as with plays from Shakespeare's First Folio, the play comes from a collected volume, transcribed as a single textfile in TCP, we give a 'playfile' number, with a numerical suffix distinguishing the play). These texts are not 'edited' or 'good' texts in the senses literary scholars are used to. TCP transcribers were told to leave blanks where they could not read the images they worked from, so there are gaps in the text (though none in these examples). Our automatic modernisation process does not get everything right ('doe' = 'do' for example, is left untreated on the mistaken assumption that it is 'doe' = female deer). It is possible to train VARD, so that errors are reduced, but for this exploratory study we decided to use the basic VARD settings so that all texts went through the same, replicable, process. We are working at scale here, and our belief is that the frequencies of the items we are counting means that 'dirt' in the data does not affect the overall result. Scholars undertaking a more focussed study might want to fine-tune the modernisation process.

A major issue to be addressed by scholars working in this hybrid field is the question of referencing within the TCP texts. As of now, the only stable reference point is the TCP file number: beyond this, there are no stable, fixed points (such as act or scene divisions, or line numbers), since these are often removed or changed in processing. For this reason, we have released the .txt files of our plays, which allows word-based searching to recover the quotations we use (see note 3).

The negative group consists of the following linguistic features or Language Action Types (LATs, using the terminology employed by our linguistic software, Docuscope):

'Standards Negative' – words and phrases indicating standards or values most people would treat negatively: e.g. *disease, unworthy, oppressed, malady, shame, the poor, unkind, envy, treason*.

'Fear' – words and phrases referencing or evoking fear: e.g. *fear, threatening, anxiety, terror, apprehension, dangerously.*

'Negative Relation' – words and phrases used to represent relationships between people which are either negatively viewed or unstable: e.g. *at war with, has been offensive, broke his heart, the rift between, draw blood*.
'Anger' – words and phrases referencing or evoking anger: e.g. *angry, vengeance, slaughter, contempt, rage, cannot forgive, coward, reproach, indignant, cruel*.
'Negativity' – words and phrases indicating negativity and negative emotions: e.g. *gloom, distrust, abhor, wretched, disappointment, warning, death, ugly, villain*.

7 The oral group consists of the following linguistic features or Language Action Types (LATs, using the terminology employed by our linguistic software, Docuscope):
'Direct Address' – pronouns and discourse markers aimed at an interlocutor: e.g. *you, you are, prithee, thy, thou, yourself, my Lord*.
'First Person' – bare first person pronouns: e.g. *I, me, mine, myself*.
'Question' – *wh-* question words and punctuation implying questions: e.g. *Who, What, Why, ?*.
'Oral Cues' – discourse markers typical of flowing speech: e.g. *nay, well, good morrow, ho, yea, ha*.
'Person Property' – words and phrases designating occupational and social roles: e.g. *brother, bondsman, mother, generals, men, sir, fellow, attendants, wife, women*.
'Imperative'- this feature identifies imperatives by looking for the base form of a verb occurring immediately after a full stop: e.g. *'. Speak,' '. Go,' '. Let,' '. Give,' '. Swear'*.

8 The self group consists of the following linguistic features or Language Action Types (LATs, using the terminology employed by our linguistic software, Docuscope):
'Self-Disclosure' - first person pronouns occurring with verbs expressing thought or consciousness (e.g. *I think, I am, I feel, I believe, I confess*) and certain pronoun-preposition combinations that function similarly (e.g. *to me, for me*).
'Autobiography' - first person pronouns used with a certain set of verbs, often past-tense, and a set of nouns that indicate past or familial relationships, in an

attempt to capture self-revelation that is rooted in a sense of the past: e.g. *I have been, I was, when I, my name, my daughter*.

'Metadiscourse' - explicit signposts from a speaker or writer to guide the audience through a piece of language: e.g. *too, we shall, but there is, either, further, moreover, aforesaid, as it were*. In our analysis, this LAT patterns with the 'self' features because it includes phrases such as 'I come', 'we come', 'I shall', 'we shall', 'I will consider', 'I will do' .

9 It may be that our 'self' features are missing some crucial marker which the tragedies associated with the 'self' in literary-critical work use to effect its linguistic representation. In which case, its identification is a job for literary criticism.

10 It is also important to point out the effect of corpus size on what we can 'see' in this type of analysis. In the context of 554 plays, the differences between Shakespeare's genres are not big enough to separate them: the similarities are more significant, and all of his plays group together. Reduce the corpus size, for example to just Shakespeare, as in our early work, and the generic differences become visible. No view is 'truer' than any other: but different perspectives allow you to see different aspects of the data.

11 The world-space group consists of the following linguistic features or Language Action Types (LATs, using the terminology employed by our linguistic software, Docuscope):

'Sense Objects' - concrete nouns: e.g. *lump, the fruit, forest, pawn, tongue*.

'Sense Property' - properties of nouns: e.g. *the appearance of, loud, hollow, round, old, voice of, sweet, hungry*.

'Spatial Relation' - words and phrases indicating location in space: e.g. *alcove, next door to, with whom, in the country, above, at the, dwell, alone*.

'Motions' - language indicating motion: e.g. *knocking, convulsions, unloose, till the, bowing, rising from, tremble, walk into the*. Much of this has a figurative emotional sense (e.g. *shudder, moved*).

12 A reminder that our results, though we present them as being 'about' Early Modern drama, are restricted by (a) what Docuscope counts; and (b) our corpus. In the future, Docuscope will be adaptable by users so that what it counts can be adjusted, or changed completely – and of course, other text analysis tools are available. The Early Modern drama corpus will be refined over time, in terms of the plays it contains, the quality of the texts of those plays (as correction projects are funded), and the quality of the metadata.

13 The language theory underpinning Docuscope, and the categories it sets up, are detailed in David Kaufer, Suguru Ishizaki, Brian Butler, Jeff Collins, *The Power of Words: Unveiling the Speaker and Writer's Hidden Craft,* London, Routledge, 2004. A number of studies illustrating its use in the classroom, and authorship work are listed at http://wiki.mla.org/index.php/Docuscope (accessed 8.5.2015).

14 Normalisation means that the raw totals for each LAT in each play are adjusted to show frequency per a set amount of words. This allows us to compare LAT frequencies between plays of different lengths.

15 The 41 LATs excluded from our study are as follows: Attack_Citation; Authoritative_Citation; CommunicatorRole; ConfirmExperience;; ConfirmedThought; Confront; Consequence; Contested_Citation; Definition; DialogCues; DirectReasoning; Example; Feedback; FollowUp; Future_in_Past; In_Media; Innovations; MatureProcess; MoveBody; NegFeedback; Neg_Citation; Negative_Attribution; PosFeedback; Positive_Attribution; Precedent_Defending; Precedent_Setting; PriorKnowledge; Procedures; Promise; Quotation; ReceivedPOV; Reinforce; Repair_Citation; Request; Responsibility; SelfReluctance; Self_Promise; Speculative_Citation; Substitution; Support; TimeDate.

16 We give a fuller account of PCA in Anupam Basu, Jonathan Hope, and Michael Witmore, forthcoming, 'The Professional and Linguistic Communities of Early Modern Dramatists' in Roger D. Sell, Anthony W. Johnson and Helen Wilcox

(eds), *Community-Making in Early Stuart Theatres: Stage and Audience* (Ashgate). Most standard statistics textbooks cover PCA (and Factor Analysis, to which it is closely related). We have found Andy Field, *Discovering Statistics Using IBM SPSS Statistics: And Sex and Drugs and Rock and Roll* (London: 2013, 4th ed.) useful. Literary scholars will probably get most out of Mick Alt, *Exploring Hyperspace: A Non-Mathematical Explanation of Multivariate Analysis* (London: 1990), which is a brief and very clear conceptual account of what the statistical procedures are trying to achieve.

17 PCA is designed to extract patterns from very complex data sets. One consequence of this is that the patterns it extracts can themselves be very complex, and difficult for human interpreters to make sense of because they consist of multiple relationships between variables – or LATs in this case. In this study, we have checked the individual distributions of the LATs we focus on in tragedies against their distribution in the drama as a whole. The box plots for this can be seen at http://winedarksea.org/?p=2013 (accessed 8.5.2015).

18 Ted Underwood, 2013, *Why Literary Periods Mattered: Historical Contrast and the Prestige of English Studies* (Stanford) – especially chapter 6, 'Digital Humanities and the Future of Literary History', pp. 157-75 – on the strange commitment to discontinuity in literary studies, and the tendency of digital/at scale work to dissolve this into a picture of gradualism. Underwood notes the extent to which this is a Romantic and post-Romantic mind-set: Classical and Renaissance approaches to literary history were very different, generally assuming genres to be trans-historical, with 'good' writers fulfilling, rather than revolutionising, generic expectations.

19 See, for example, on the frequency of 'the' in *Macbeth*, Jonathan Hope and Michael Witmore, "The Language of *Macbeth*", chapter in *Macbeth: The State of Play*, edited by Ann Thompson, London, Bloomsbury (Arden), 2014, pp. 183-208.

20 For a discussion of 'influence', see 'What is Influence?' http://winedarksea.org/?p=1629 with comments from Matt Jockers and Ted

Underwood (accessed 8.5.2015), and Bill Benzon's detailed reading of Jockers' *Macroanalysis*, the full version of which can be downloaded from this URL: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2491205 (accessed 8.5.2015).

21 Alan Farmer is currently working on estimates of the number of texts and editions we have lost. Although we can never know for sure how much material has not survived, it will be important for future users of EEBO-TCP to remind themselves that, however large the digitised corpus, it is always incomplete.

Further Reading

Mick Alt, 1990, *Exploring Hyperspace: A Non-Mathematical Explanation of Multivariate Analysis* (London: McGraw Hill)

Jonathan Hope and Michael Witmore, 2014, "The Language of *Macbeth*", chapter in *Macbeth: The State of Play*, edited by Ann Thompson, (London: Arden), pp. 183-208

Russ McDonald, 2006, 'The language of tragedy', in Claire McEachern (ed.), *The Cambridge Companion to Shakespearean Tragedy* (2nd ed., CUP), pp. 23-49