Networks and Communities in the Early Modern Theatre

Anupam Basu, Washington University in Saint Louis, USA
Jonathan Hope, University of Strathclyde, Glasgow, UK
Michael Witmore, Folger Shakespeare Library, Washington DC, USA

(Research for this chapter was funded in part by the Mellon Foundation, and The Royal Society for Edinburgh)

This chapter presents initial findings from a larger and on-going digital project, *Visualising English Print*.[1] Here, we seek to demonstrate the potential of emerging computational techniques that make use of databases, linguistic taggers, network analysis and visualization for extending our understanding of the dynamics of relationships in the early modern theatre. We highlight evidence for the existence and nature of communities defined by professional relationship within the institutions of the theatre, and also for communities formed by literary relationship as evidenced by linguistic similarity and difference.

Much of this work deals with experimental tools and techniques and remains at an early stage. The data sets we are using are impressive in many ways, but are not perfect, either in terms of coverage, or internal organization. We present this work as an indication of what is likely to be possible in the near future, and as part of the process of testing and refining our data and techniques. The accounts of Early Modern theatrical communities we present (which we might call hypotheses) are likely to change as we continue our work and as other scholars work with the same tools and data sets.

Communities of Professional Relationship

Thomas Heywood claimed in 1633 that he had had 'either an entire hand, or at the least a maine finger'

[1] See the project website http://graphics.cs.wisc.edu/VEPsite/ (accessed 12.2.2014) and project members' blog http://winedarksea.org (accessed 12.2.2014).

in the composition of 220 plays.[2] Even if exaggerated, his claim points to the intensely collaborative nature of theatrical production in early modern London, a process that was as much economic division of labor as artistic collaboration. But any student of early modern drama who has pored over editorial notes on the style and habits of individual contributors to a play will recognize that even in this strangely ad-hoc conveyor belt of cultural production that churned out plays to meet the insatiable demands of the London audience, there were genuine spaces of artistic collaboration.

We will present a series of visualisations exploring the collaborative networks that emerged in the early modern theatre. The data underlying these visualisations is extracted from DEEP (Database of Early English Playbooks), a database hosted at the University of Pennsylvania that collects information on every playbook printed in English up to 1660.[3] The database draws on a wide variety of scholarly sources on the early modern stage including, among others, Harbage, Gregg, and Bentley in a format that allows users to perform various types of searches and extract various subsets of the data.[4] However, while such a resource provides scholars with an incredibly powerful and versatile exploratory tool, using visualization and network analysis techniques can allow us to get a sense of the data in ways that would be impossible (or very difficult) from the tabular representation alone. It might also enable us to perceive relationships between individuals and entities on multiple levels, and at the same time throw into stark relief some of the ways in which the imposition of a tabular structure that such a database implies can be at odds with the complex, and contested narrative of the emergence and development of the early English theatre.

---

[2] Thomas Heywood, 1633, 'To the Reader' in *The English Traveller*, A3r (quoted from JISC Historic Books, page opening 3, open URL: http://www.jischistoricbooks.ac.uk/Search/?bibnumber=STC%20(2nd%20ed.)%2013315.&spage=3 (accessed 12.2.2014)

3 Lesser Zachery and Alan Farmer, "DEEP: Database of Early English Playbooks," accessed January 27, 2014, http://deep.sas.upenn.edu/.

4 For a complete listing of sources, see http://deep.sas.upenn.edu/sources.html

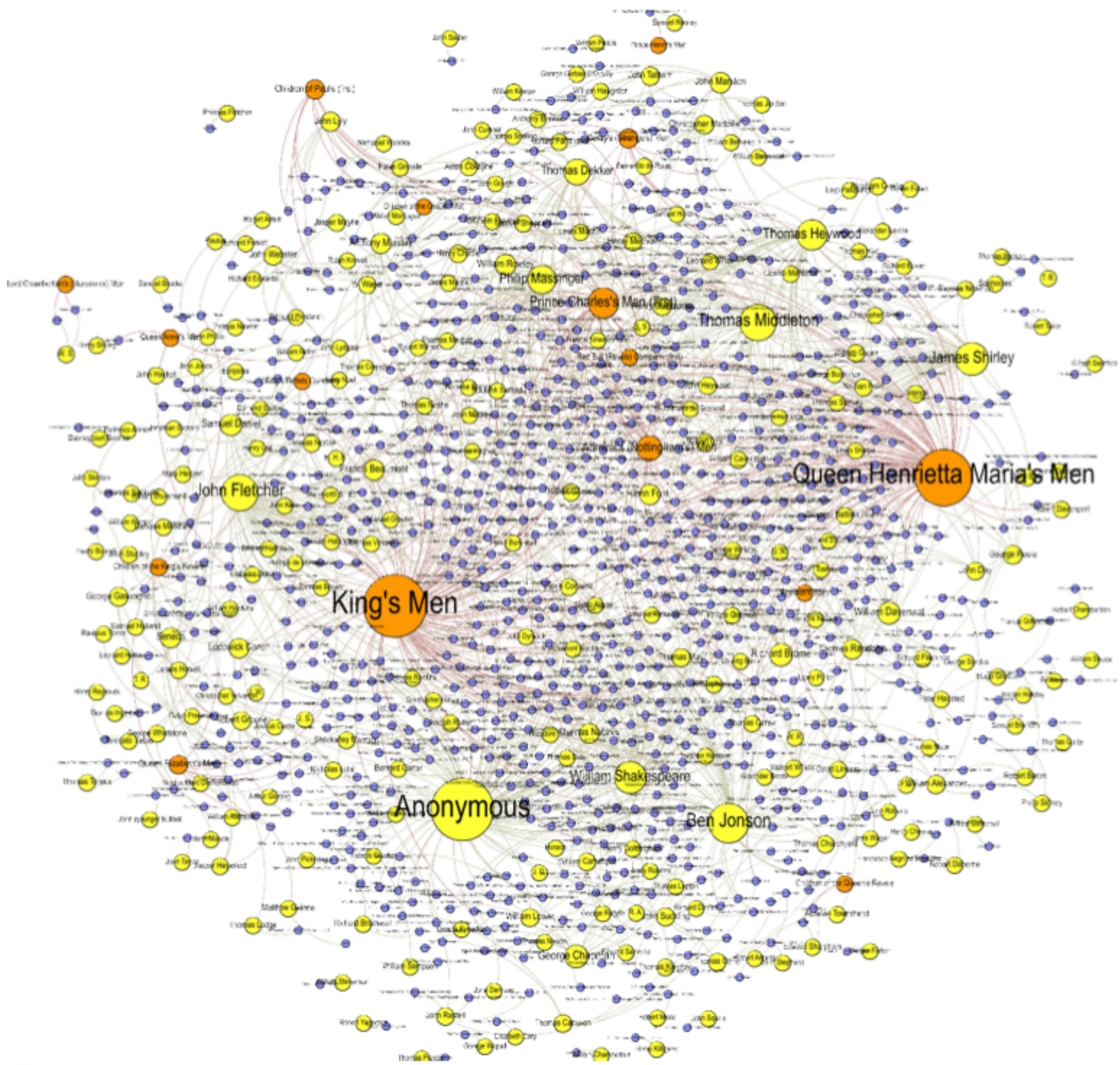


Illustration 1: Fig. 1

We begin with a network that attempts to capture a sense of the interlinkages and collaborative spaces of the early modern theater for a period spanning roughly a century (Figure 1). The network itself is too busy to be readable without a zoomable format, but it does capture the effect of a theatrical community bustling with commercial activity. We represent plays (blue), authors (yellow) and companies (orange) as 'nodes' in this network and lines between them represent connections or, in network terminology, 'edges'. Complex as this network looks, with over a thousand nodes, it represents a striking simplification of the contested history of development of the early modern theater. It doesn't visualize time very well, for example, or the evolving chronologies of the theater companies, only gesturing at these in the way various major nodes organize themselves. More importantly, and perhaps more insidiously, the nature of our data privileges print over performance, making the playbook (and the accidents of its survival) our window into the socio-economic world of the theater while the irreducible gaps in our knowledge of the early modern theater leaves its trace in the form of the prominent 'author' node labeled 'anonymous'. But the evolution of the English theater – from ritual performance or uneasy adaptation to the most glorious literary productions of the European renaissance – leaves its mark in these network visualizations as much in the form of telling absences and simplifications as in the sense of teeming productivity that attests to its commercial and creative vitality.

Thus, even as it captures a sense of the complexity of the relationships and communities that constituted the early modern theater in England, such a visualization invites us to reflect on the processes of computational analysis and the nature of the underlying data as well. Any visualization is not only a representation of data but also an interpretation of it. And like any complex interpretation, it contains its ambiguities and aporias, its contested narratives and hidden subtexts. It can throw certain features into stark relief and de-emphasize or elide others. This is not entirely surprising as most visualizations seek to foreground particular aspects of the data, often at the cost of others, but perhaps it bears repeating in a situation such as this, where emergent technologies that are brought to bear on nuanced cultural phenomena might seem to imply a determinism that is alien to the way we encounter cultural history. Indeed, this question of the interpretive status of visualisation and its relation to the underlying data is one we will need to repeatedly encounter as we refine our initial network diagram to bring out particular collaborative networks.

Illustration 2: Fig. 2

Our second network extracts company affiliations for individual playwrights (Figure 2). The major companies figure as large, light circles, while the writers are smaller dark ones. A professional relationship is represented by a line. Although we need to treat this visualization with particular care, since DEEP does not always record subsequent relationships (and they are not all represented here), it should be clear that the picture of the Early Modern theatre, and its communities, represented here is rather different. Where Figure 1 presented us with a rather diffuse network of multiple links, here we have a more dense network, with a set of very clear nodes: the acting companies. On this representation at least, the Early Modern theatre is highly organized around acting company, with most writers restricted to working for one, and a very small number of writers – Massinger, Rowley stand out here – having multiple affiliations.

It is crucial to interrogate the particular biases and blindnesses of the database as the underlying construct that creates this sense of structure. No data-set is perfect, and in fact the constraints that the tabular form of a database imposes can be at odds with the complex narrative history of the early theatre. In this case, impressive as DEEP is, we know that it (and, at times, the underlying sources that DEEP uses) does not record all the complexities of the shifting theatrical relationships in Early Modern London. In following Bentley's convention of collapsing multiple stages in the evolution of theatrical companies into one single label, it elides the fragmented nature of development of most of these companies. Partly, the network seems reductive because the companies themselves were far less stable than the entries in the database suggest. 'The King's Men', for example, is an umbrella term for a set of companies spanning over more than half a century, which were only loosely related to each other. The traditional emphasis on aristocratic affiliation might not be the best way to think of the evolution of the theater. That might be a narrative best told with data centred on playhouses, entrepreneurs, major actors and playwrights – data that might give us a very different perspective than this relatively neatly (perhaps misleadingly) organized network of affiliations. Again, many things are missing from this visualisation: time and individual plays (which would give us an idea of the extent of each relationship) to name just two. But more information tends to make visualisations harder to 'read', as we can demonstrate by extending Figure 2 into Figure 3, which presents the same data, but retains individual texts as nodes mediating the connection between authors and companies.

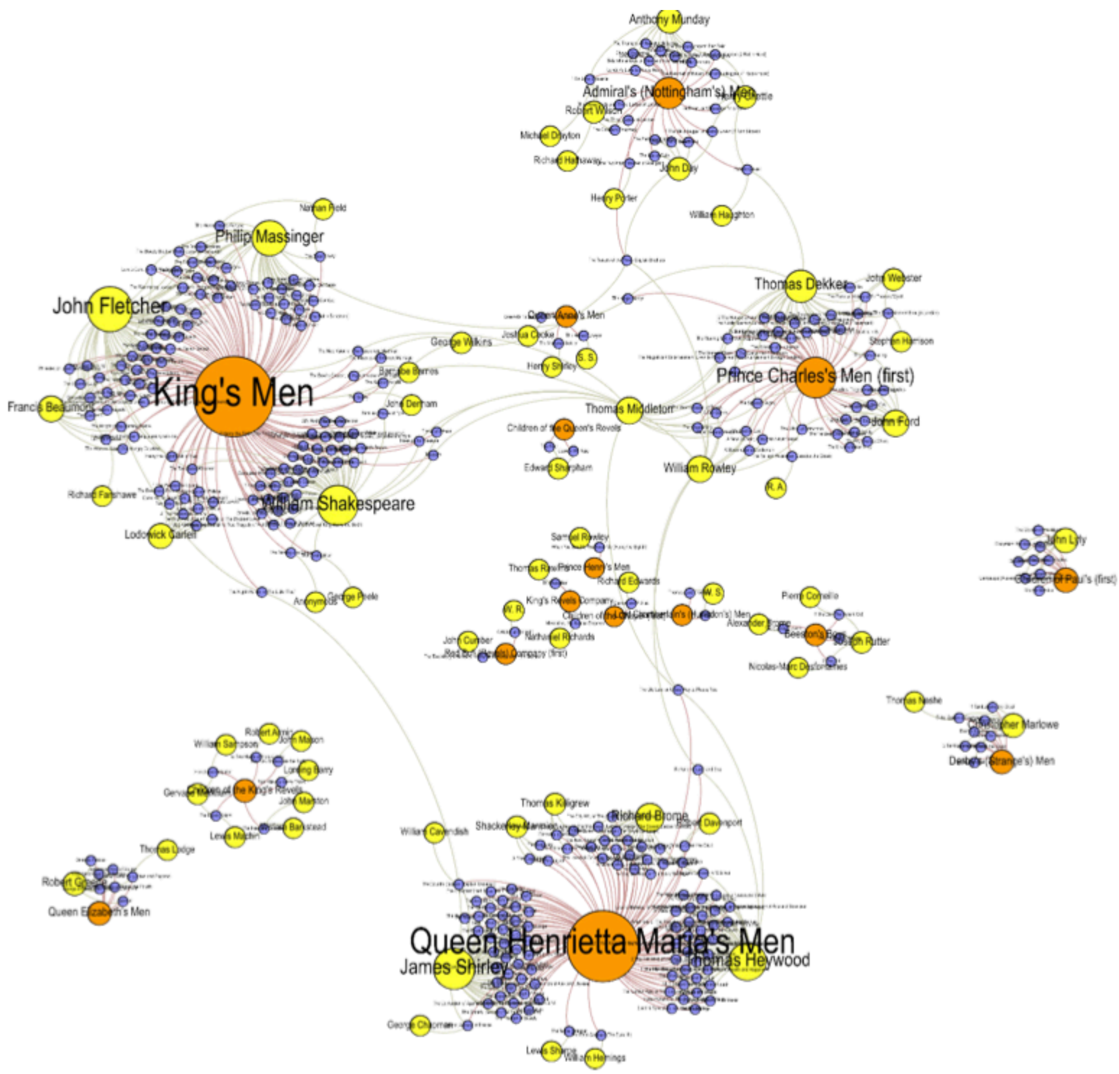


Illustration 3: Fig. 3

Figure 3 is a 'tri-partite' network, in that its nodes represent the relationships between three different kinds of entities – the playbook, the theatre company, and the individual playwrights. While it is obvious that adding the individual plays does not fundamentally change the structure of the network, making individual plays visible does give a strong sense of the degree of association between plays and companies, and also of the companies' repertoires. Individual plays thus perform a vital connecting element in extracting collaborative networks that we can use to our advantage to generate a network of collaborations.

In figure 4, we extract only the play-nodes that are attached to more than one author-node. In other words, we visualize only those plays that are marked as having multiple authors in the database. It should be noted that this does not quite map neatly to authorial collaborations because the logistical decisions made in the construction of the database blur the lines between actual collaboration and translation. We notice many continental and classical authors appear to be part of this network. Seneca, for example, puts in a strong showing on 'collaborations' with Elizabethan dramatists. Absurd as it might seem, this is in a sense quite appropriate, as it maps intellectual affiliations rather than mere spatial, economic, or chronological affinity. We might also notice that many of these collaborations form isolated islands (or, disjoint communities, to use Social Network Analysis terminology). These indicate mostly coterie drama written by occasional writers who translated or adapted classical works in many cases. The dataset does 'clean' up many of the ambiguities associated with early modern print culture by noting the modern scholarly consensus on many anonymous or misattributed authorship claims, drawing on various sources. But while we must keep these caveats in mind as we study this network, these are mostly artifacts of the way the underlying data is organized rather than the computational process that creates the visualisation. Whether or not to categorize translations or adaptations as collaboration rather than a distinct type of literary encounter is an interpretive decision that the editors of DEEP and, in some cases, the editors of their scholarly sources make, and it is one that is thrown into relief by the network diagram. In fact, one of the salutary effects of such computational interventions in literary datasets is to throw light on such possible anomalies or instances where one would want more nuance in the way we record data. It is quite possible, from a technological perspective, to either revise the underlying database or control for certain exceptions before regenerating the network, but in this instance we present the information from DEEP in unmodified form to see to what extent we can deduce the traces of collaboration from it.[5]

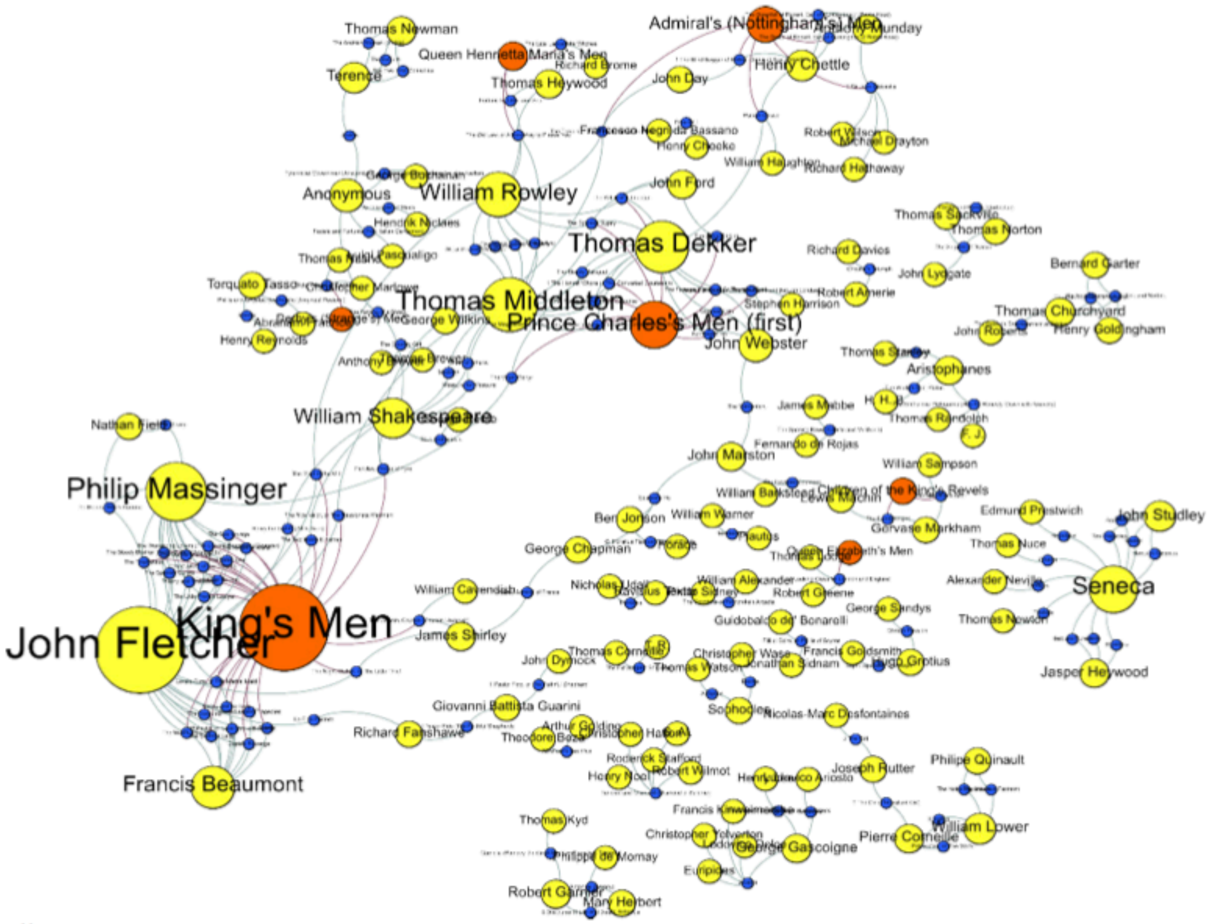


Illustration 4: Fig. 4

Since we are now interested mainly in authorial collaborations, rather than company affiliations, what happens when we simplify the plot to indicate only links between playwrights, eliminating the companies and plays as nodes? Figure 5 captures the linkages between playwrights as direct edges and

[5] Eventually we would expect DEEP, and other sources of metadata on the Early Modern theatre, to be updated from the huge revision of such knowledge currently being undertaken by Martin Wiggins and Catherine Richardson in *British Drama 1533-1642: A Catalogue* (currently 3 vols.; Oxford: 2011 – ongoing).

Figure 6 extracts the single largest connected island of nodes from this network. The smaller islands that are eliminated represent coterie collaborations by dramatists who never played a major part in the commercial theatre of early modern England, but the central connected community that emerges from this network presents a more familiar picture, even if one that must be read with a set of strong caveats. This visualization takes each of the major early modern playwrights, and shows his collaborations via a single line to each collaborator. The lines are weighted by the number of shared collaborations and the size of the node for each playwright corresponds to the number of collaborative plays attributed to him in the database. These playwrights form the core of the early modern canon and wrote for the commercial stage, often collaborating repeatedly with a given partner. Fletcher, Middleton, Rowley, Beaumont, Dekker and Shakespeare form the core of this collaborative group. Of the major playwrights, Webster has relatively little connection to the group but most conspicuous by his near-absence is Jonson, who remains only very tenuously connected to the network. Jonson's insistence on excising collaborations from his published work contributes to this, another artifact of privileging print editions over the many other sources that we might draw on to map the early modern theater. Alone amongst the 'major' names, he collaborates with only two others. Even Shakespeare, who certainly collaborates with relatively fewer people than the central group, and arguably collaborates less extensively, looks more 'normal' in terms of his collaboration than Jonson.

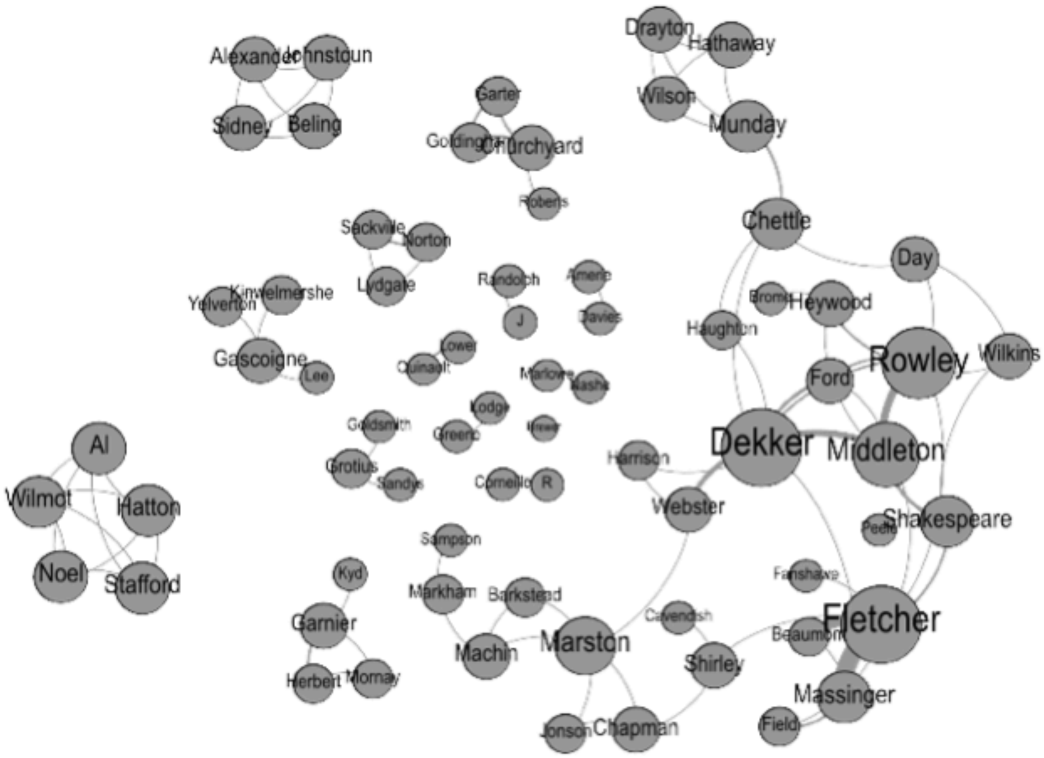


Illustration 5: Fig. 5

Illustration 6: Fig. 6

Time, always difficult to accommodate in network diagrams, might have added a crucial dimension to the visualization of such a dataset. It might have been especially significant in Shakespeare's case, where the sense is that he collaborates early, and late in his career, but less so in the middle years. The elimination of individual plays comes at a slight cost since now the information about the degree of collaboration needs to be compressed into the relatively harder to read visual aspects of the sizes of the nodes and the thickness of the links connecting them. And in eliminating authors not connected to the central network, we have consciously decided to emphasize the commercial theatre as the main site of collaborative community formation in early modern England.

These visualizations, like much digital analysis, do not present us with answers or final accounts. Rather, they provoke questions and possible further research: to what extent do these representations chime with our current picture of the early modern theatrical world? Are the communities they suggest, centered on theatrical company, 'real'? Would the inclusion of time in some way allow us to picture the set of relationships more usefully? In what way can we account for the incongruities and silences that are thrown into relief by such computational intervention? Which of these artifacts are easily corrected for as requiring quite simply a 'better dataset' (perhaps one that distinguishes between actual collaborators and translators) and which point to deeper implications of the ways in which we tend to privilege print culture as a gauge of theatrical communities (a database trying to counter this impulse would account better for Jonson's collaborations even though they were excised from print in many instances). 'Better data or better algorithms' has been the central point of contestation in many debates on computational theory.[6] In the case of literary analysis, we'd suggest we need to strive for both. These tools are new and offer a radically unfamiliar perspective on the way we construct and think about literary and cultural history.[7] As we seek to improve on and account for their weaknesses, hopefully we will be able to gain new insights from some of their strengths.

Communities of Linguistic Practice

So far in this chapter, we have used established 'facts' about professional relationships in the Early Modern Theatre as a basis for visualizations which attempt to make patterns within those relationships

[6] Compare, for example http://allthingsd.com/20121128/better-algorithms-beat-more-data-and-heres-why/ and http://anand.typepad.com/datawocky/2008/03/more-data-usual.html (both accessed 12.2.2014).
[7] Ted Underwood, in a number of blog posts, has outlined this challenge to literary scholars. See, for example http://tedunderwood.com/2013/02/08/we-dont-already-know-the-broad-outlines-of-literary-history/ (accessed 12.2.2014).

clear to us. We have emphasized two key lessons to be borne in mind in any digital work:

(1) visualizations have the appearance of solidity and 'truth' (they are inherently convincing in a way prose descriptions of the same data are not: visual rhetoric trumps verbal): a line drawn in a network looks 'real', but it is simply a representation of a piece of information in a matrix – which in this case is a representation of a 'fact' in DEEP, itself a record of a claim in the standard literature on the Early Modern Theatre – the 'truth' of a visualization rests, not in the visualization software, but in the data that lies behind it

(2) visualizations generally work best by *leaving out* data: Figure 1 has too much data for us to make sense of it; the data reductions we present in Figures 2 and 3 give a clearer story, and are explicable in ways Figure 1 is not, but this clarity is achieved at the expense of much (if not most) of the available data. The same is true of almost all data analysis: to get at the story of the data, we need to select. This may seem counter-intuitive, but it is directly parallel to what is normal practice in literary criticism: all literary critical accounts of the Early Modern theatre are explicitly or implicitly selective, leaving out writers, texts, periods, genres for the sake of producing a coherent narrative.

In this section of the chapter, we shift from analyzing information about professional relationships, to the analysis of linguistic behavior. The basic question we seek to answer is 'Are the communities of professional relationship identified in part 1 also visible as communities of linguistic behavior?'. Put another way: 'Do Theatre Companies, or any other groupings of writers, have recognizable linguistic styles?'. In order to answer this question, we need to prepare an electronically searchable corpus of Early Modern plays. We then need to count linguistic features we think contribute to 'style'.[8] We then

[8] 'Style' is, of course, a highly complex term, with different (sometimes contradictory) meanings between fields, and even within the same field. The decision of what to count in order to measure

need to compare the various frequency counts for each linguistic feature of the plays to see which groups of plays show similar 'styles', and which groups differ. Our working hypothesis is that writers in professional communities will show linguistic similarities with each other, and these will be differentiated from those of other communities (the tendency for individuals in well-defined communities to develop coherent linguistic behaviours that mark those communities is well-attested in sociolinguistics and historical-sociolinguistics[9]). This may be evident in characteristic 'house styles' belonging to the main theatre companies, or in stylistic continuities observable in other groupings of writers.

Our linguistic analysis is carried out on a new data set. Rather than the metadata contained in DEEP, we use a collection of 591 transcribed play texts from the Early Modern period.[10] To do the analysis we have used a program called Docuscope, which Hope and Witmore have used extensively in the past.[11] We ran all 591 plays through Docuscope, which counts the frequencies of 101 linguistic features known as Language Action Types (LATs), and used the statistical software package JMP to analyse the

---

style is a subjective, interpretive one which must be made by the investigator. It is a problem for literary scholars, not statisticians or digital experts.

[9] See, for example, Suzanne Romaine, 1982, *Socio-historical Linguistics: Its Status and Methodology* (Cambridge) and Terttu Nevalainen and Helena Raumolin-Brunberg, 2003, *Historical Sociolinguistics: Language Change in Tudor and Stuart England* (London).

[10] The texts we use come from the EEBO-TCP transcriptions, and were 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 from copies held by the Folger Shakespeare Library. Texts were modernized automatically, using VARD (http://ucrel.lancs.ac.uk/vard/about/ - accessed 12.2.2014) and were subject to some minor clean-up to remove certain characters introduced during transcription. Texts were stripped of all non-spoken elements (stage directions, act and scene numbers, speaker designations).

[11] See Jonathan Hope and Michael Witmore, "The very large textual object: a prosthetic reading of Shakespeare", *Early Modern Literary Studies,* 9.3, Special Issue 12, January 2004, 6.1-36 http://www.shu.ac.uk/emls/09-3/hopewhit.htm (accessed 12.2.2014); Michael Witmore and Jonathan Hope, "Shakespeare by the Numbers: On the Linguistic Texture of the Late Plays", in Subha Mukherji and Raphael Lynne, eds, *Early Modern Tragicomedy*, Cambridge, D.S. Brewer, 2007, pp. 133-153; Jonathan Hope and Michael Witmore, "The hundredth psalm to the tune of 'Green Sleeves': Digital Approaches to the Language of Genre", *Shakespeare Quarterly*, 61, no. 3, Fall 2010, pp. 357-90 http://muse.jhu.edu/journals/shakespeare_quarterly/toc/shq.61.3.html (accessed 12.2.2014).

resulting spreadsheet file of frequency counts.[12] We will discuss LATs, with examples, once we begin the detailed analysis of the results; for now, we discuss the process of visualizing and analyzing the data produced by counting anything in a corpus.

When we count an attribute of a number of things (for example, the frequency of a certain linguistic feature in a group of plays), we effectively plot the plays in a data space. This sounds complex, but is (to begin with) quite straightforward and easy to understand. For example, one of the LATs we count with Docuscope is called 'Oral Cues'. This is a LAT designed to capture linguistic features which convey the impression of actual speech (it tags words such as 'well', 'my word', 'good morrow', 'ah', 'Yes sir', 'ye', 'No,' 'nay', 'ha', and so forth). Docuscope counts all of the words or phrases it recognizes as belonging to the category 'Oral Cues' in each play, and standardizes the raw frequency count for each play by dividing it by the total number of tagged words in that text, and multiplying by 100 to give a percentage. This enables us to compare between texts of different lengths. Docuscope then outputs in a spreadsheet the percentage of words tagged in each play as 'Oral Cues'.

If we 'sort' the spreadsheet on the column for 'Oral Cues', the software will put the plays in order of their relative use of 'Oral Cues'. Table 1 is taken from the spreadsheet of Docuscope results for all 74 LATs over 591 plays.[13] It shows the ten least frequent plays in terms of their use of this LAT, and the ten most frequent (omitting 571 plays in between). The first column gives the unique play code for each play in the corpus, the next gives the author, the next the title, the next the date[14], the final column

---

[12] www.jmp.com (accessed 12.2.2014).

[13] Although Docuscope counts frequencies for 101 LATs, we exclude from our analysis any LATs which have scores of zero in any plays. This leaves 74 LATs for this data set.

[14] The metadata we have for this corpus was extracted by TCP transcribers and is an amalgamation of metadata from EEBO and ESTC. Like DEEP (see footnote 4), these sources are in turn based on the standard sources of information about the Early Modern theatre, and title page information. In many cases, this metadata has been superceded by subsequent scholarship: the process of cleaning it up will be a long one. It is important for literary scholars working with data sets such as this to get used to the uncomfortable notion that the metadata, like the texts, is imperfect, and probably always will be.

gives the percentage score for the LAT 'Oral Cues'. Thus the play in our corpus with the lowest use of 'Oral Cues' is Walter Montagu's *The Shepherd's Paradise*, where only 0.041978% of the tagged phrases are from this LAT. The play with the highest use of 'Oral Cues' is John Fletcher's *The Wild Goose Chase* (1625), where 1.64345229% of tagged items are from this LAT (this may seem a very low figure, but to give some sense of how frequent the LAT is in the play, the raw score for 'Oral Cues' in this text is 135 items – the raw score of *The Shepherd's Paradise* is 35).[15]

Table 1: 'Oral Cues' in the 591 play corpus: lowest and highest ten plays by relative frequency

playfile	author	title	date	OralCues
A07649	Montagu, Walter	The Shepherd's Paradise	1659	0.041978
A02262	Grotius, Hugo	Christ's Passion	1640	0.04504166
A02455	Habington, William	The Queen of Aragon	1640	0.06878307
A73627	anon.	Caesar and Pompey	1595	0.07799361
A07974	Nabbes, Thomas	Hannibal and Scipio	1637	0.10827068
A11909_07	Seneca, Lucius Annaeus	Medea	1581	0.10872191
A18404_01	Chapman, George	Conspiracy of Byron	1608	0.10927431
A02227	Greville, Fulke	Mustapha	1596	0.12130198
A19738	anon.	The Wars of Cyrus	1594	0.12196872
A11909_01	Seneca, Lucius Annaeus	Hercules Furens	1581	0.12856437
…				
A27177_10	Fletcher, John	The Chances	1625	1.11755082
A27177_01	Fletcher, John	The Mad Lover	1625	1.11853502
A27177_11	Fletcher, John	The Loyal Subject	1625	1.13590548
A63300	Tatham, John	The Scots Figaries	1652	1.15285942
A03424	anon.	Band, Cuff, and Ruff	1615	1.16015846
A09857	Porter, Henry	1 Angry Women…	1599	1.29946419
A18374	Chamberlain, Robert	The svvaggering damsell	1640	1.41040766
A14193	Udall, Nicolas	Ralph Roister Doister	1552	1.56868355
A27203	Fletcher, John	The Wild Goose Chase	1621	1.64150391
A27177_36	Fletcher, John	The Wild Goose Chase	1625	1.64345229

Simply by laying out the plays in order in Table 1 we have visualized the data – because the order in which the plays appear on the page represents an aspect of the data – and this visualization allows us to

This is counter-intuitive for literary scholars, who tend to want 'perfect' texts and metadata before they start work, but that represents an ideal, and impossible, situation. Working at scale, statistics tells us, means you can cope, to a degree, with messy and missing data - but we have to get used to treating metadata with critical scepticism, and knowing its limitations.

[15] Readers will note two entries for *The Wild Goose Chase* in this table: this is because our corpus contains multiple versions of plays where plays were republished in the period. It is arguable that we should exclude duplicates from our analysis, though some 'duplicates' are revisions, and arguably therefore different texts – while even direct reprints could be said to be representative of the corpus of drama at a particular time, and therefore deserving of a place in the corpus.

see (even from this attenuated table) that John Fletcher scores very highly on this LAT. And we can make this visualization even more informative by adjusting the spaces between plays in the table in proportion to their relative frequency scores – which gives us an idea of how extreme some of the results are (see Table 2). Now we can see that the plays with low 'Oral Cues' scores are tightly and evenly grouped together in terms of their score, while the plays with high scores are strung out over a greater distance with large gaps between them.

Table 2: 'Oral Cues' in the 591 play corpus: lowest and highest ten plays by relative frequency, spacing proportional to relative frequency

playfile	author	title	date	OralCues
A07649	Montagu, Walter	The Shepherd's Paradise	1659	0.042
A02262	Grotius, Hugo	Christ's Passion	1640	0.045
A02455	Habington, William	The Queen of Aragon	1640	0.069
A73627	anon.	Caesar and Pompey	1595	0.078
A07974	Nabbes, Thomas	Hannibal and Scipio	1637	0.108
A11909_07	Seneca, Lucius Annaeus	Medea	1581	0.109
A18404_01	Chapman, George	Conspiracy of Byron	1608	0.109
A02227	Greville, Fulke	Mustapha	1596	0.121
A19738	anon.	The Wars of Cyrus	1594	0.122
A11909_01	Seneca, Lucius Annaeus	Hercules Furens	1581	0.129
...				
A27177_10	Fletcher, John	The Chances	1625	1.118
A27177_01	Fletcher, John	The Mad Lover	1625	1.119
A27177_11	Fletcher, John	The Loyal Subject	1625	1.136
A63300	Tatham, John	The Scots Figaries	1652	1.153
A03424	anon.	Band, Cuff, and Ruff	1615	1.160

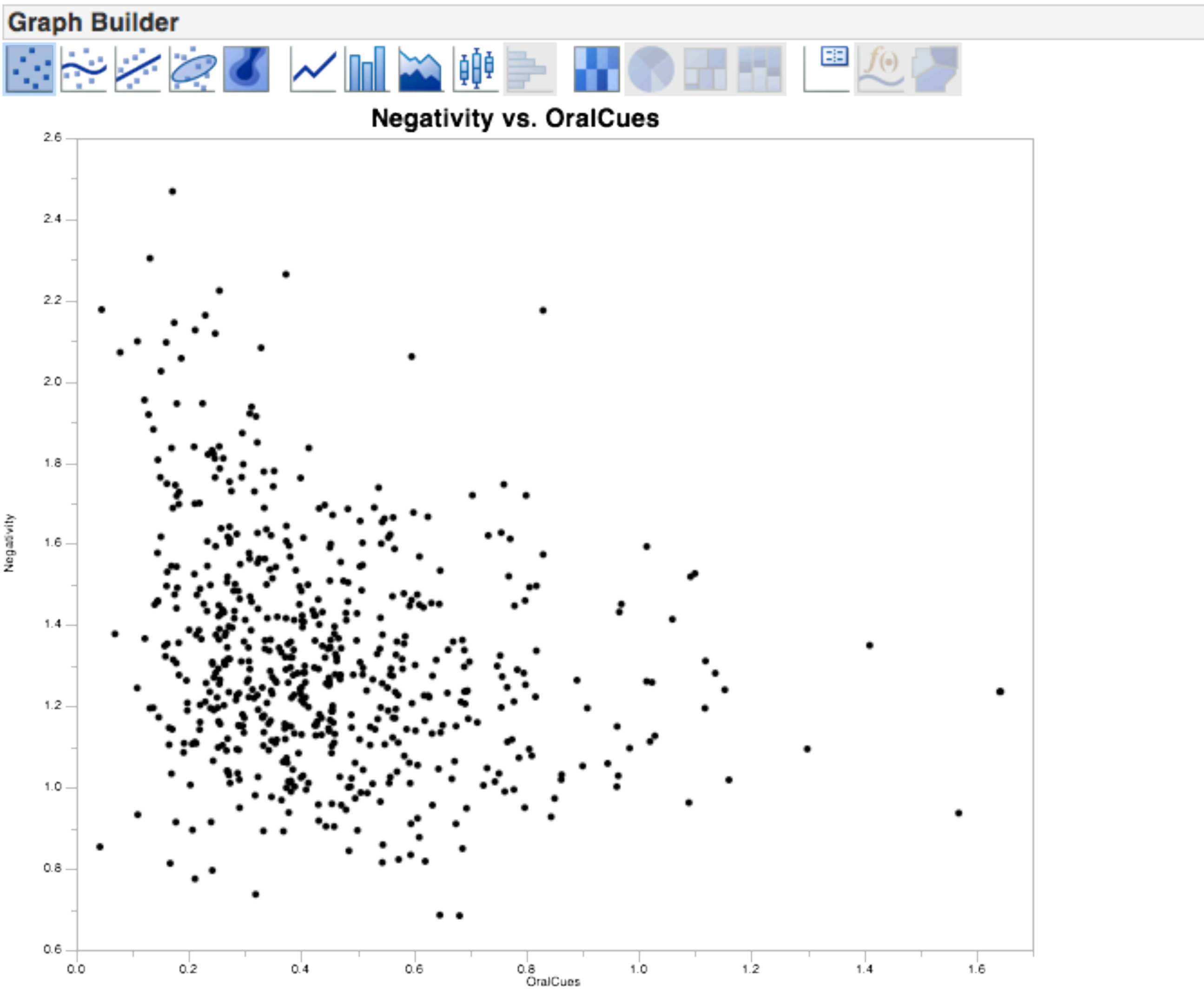
A09857	Porter, Henry	1 Angry Women…	1599	1.299
A18374	Chamberlain, Robert	The svvaggering damsell	1640	1.410
A14193	Udall, Nicolas	Ralph Roister Doister	1552	1.569
A27203	Fletcher, John	The Wild Goose Chase	1621	1.642
A27177_36	Flecther, John	The Wild Goose Chase	1625	1.643

What we have done here is project the data into one-dimensional space: ordering the plays by relative frequency of 'Oral Cues' allows us to see the relationships between them on this variable much more easily than if the spreadsheet were ordered by author, or date, or file code. But of course, we have achieved this clarity by leaving out 73 variables: a massive reduction in the data.

We could continue projecting the plays into one-dimensional space for each of the 74 LATs, laying them out in spreadsheet columns. This would allow us to see the distribution of plays on each LAT, but it would be hard for us to see relationships between LATs: Does a high score on 'Oral Cues' predict a high score for any other LATs? Or is it always accompanied by a low score on certain LATs? Intuitively, we might predict that 'Oral Cues' ought to be tracked by other LATs associated with oral, or informal, features, while LATs associated with formal features ought to pattern against it, but searching for such relationships across 74 spreadsheet columns would be difficult and slow.

Once we start projecting data into more than one dimension, however, it becomes easy to explore relationships between variables. In Table 2, by making space proportional to relative frequency, we projected the data onto a line with 0% at one end, and, theoretically, 100% at the other. If we take two variables and project them both onto separate lines, we can use those lines as the axes of a graph to project our 591 plays into two-dimensional space – which we have done in Figure 7. Here, each play has two values, one for each LAT: these become the coordinates for the play in the two-dimensional space of the graph (we have added scores for the LAT ‘Negativity’, which tags words associated with negative states and associations).

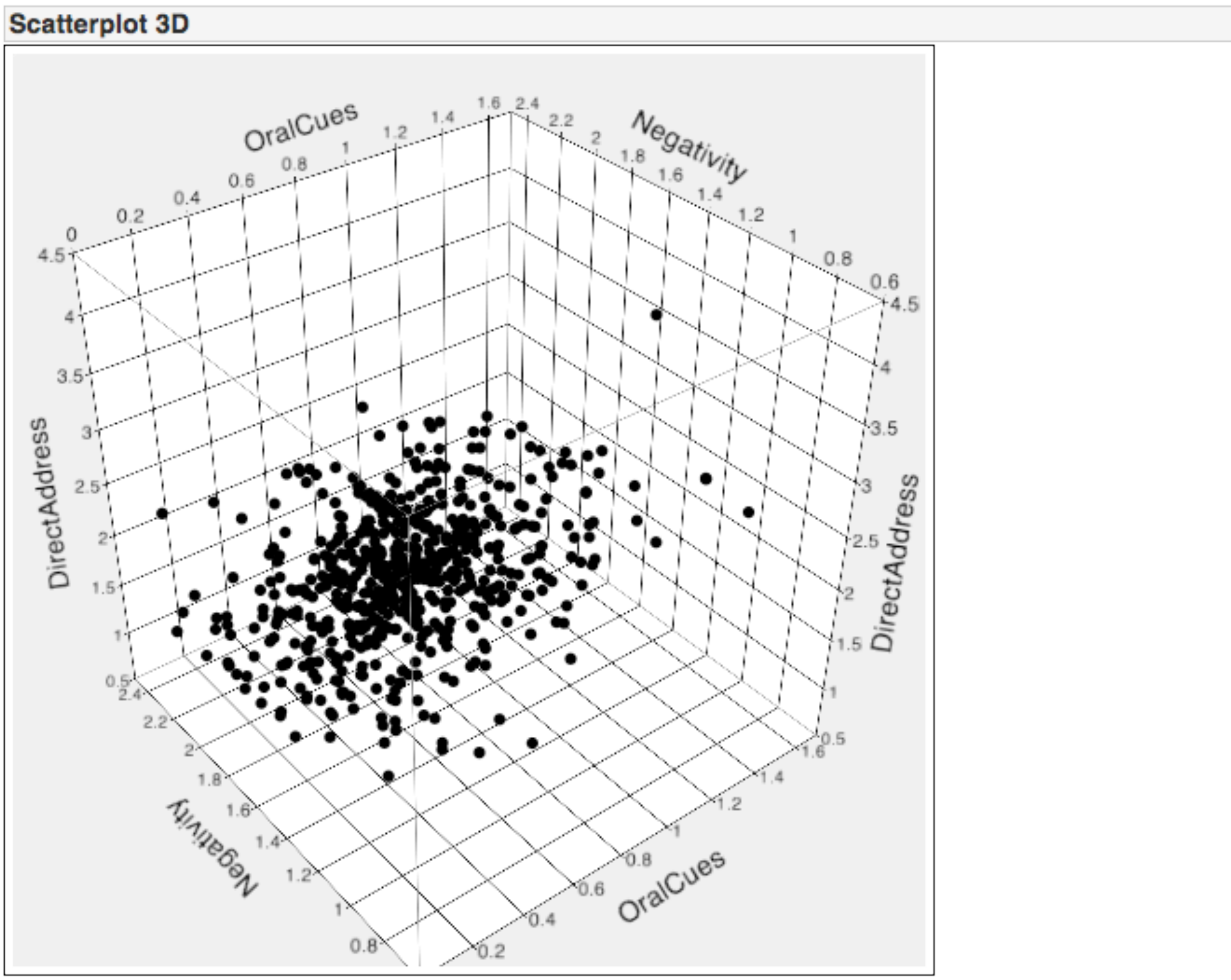
Figure 7: The 591 play corpus (black dots) plotted in two-dimensional space using two variables ('Oral Cues' and 'Negativity')



This allows us to compare the relationship between the two variables: we can see from the graph that most plays fit into the lower right, suggesting that if a play's score is average on one of these LATs, then it will also be average on the other. At the extremes, something else seems to be happening however: plays that are exceptionally high on 'Oral Cues' (to the right hand side of the graph) tend to be low to average on 'Negativity' (they appear in the lower half of the graph), while plays high on 'Negativity' (upper half of the graph) tend to be low to average on 'Oral Cues' (they appear on the left side of the graph). For example, Fletcher's *The Wild Goose Chase* (both versions) are plotted on the extreme right of the graph, at just over 1.6 on the horizontal axis – and now we can see very clearly, from its distance from other plays, how extreme the result of this play is on this LAT. But we can also see that the value of this play on the vertical axis ('Negativity') is very much in line with that of most plays (between 1.2 and 1.4). So extreme use of one LAT does not trigger extreme use of the other.

It is therefore easy to move from one to two dimensions by treating the relative frequency scores as spatial coordinates. This enables us to include more data in our visualisations, search for groups of plays that behave similarly (they occupy the same space in the visualization), and compare the interactions between variables. By adding a third variable, and treating that as a spatial coordinate too, we can project our data into three dimensional space. JMP allows us to do this, and will produce a three-dimensional animation that can be rotated to allow us to see the patterning of the plays (see Figure 8 - unfortunately the animation is not reproducible in this book!):

Figure 8: The 591 play corpus (black dots) plotted in three-dimensional space using three variables ('Oral Cues', 'Negativity', and 'Direct Address')



Indeed, the mathematical principles of adding dimensions do not change beyond three: we can continue to add variables, treating the value of each one as a spatial coordinate, right up to the total number of variables for which we have data. So in this case, we have projected the 591 plays into 74-dimensional space, with each play located at a unique point specified by a set of 74 coordinates. At this point, of course, we run up against the limits of human cognition: we can't see more than three dimensions, and we have trouble imagining more.[16] Mathematics, however, can easily model spaces which consist of more than three dimensions – so once we have projected our play corpus into a 74-dimensional space, we can measure the distances between plays in order to compare them, and get a sense of the similarity and difference between items in the hyper-dimensional spaces we have created. We just can't see those dimensions.

What we can do, however, is use various statistical methods to reduce the dimensionality so that we can visualize patterns in the data. One common way of doing this is to use principal component analysis (PCA).[17] PCA is a way of reducing the 74 dimensions we have plotted the plays onto into a set of fewer 'components', loading as much of the hyper-dimensional information on each component as possible. Imagine, if you can, our original one-dimensional graph multiplied so that instead of one axis, we have 74, all pointing in different dimensions. One way of thinking of components is that they are lines drawn in hyper-dimensional space which try to run as close as possible to as many of the existing

[16] Curiously, Early Modern drama, in the shape of Shakespeare, has a significant history in attempts to imagine hyper-dimensional worlds. E.A. Abbott, the author of *A Shakespearian Grammar* (London, 1870), also wrote *Flatland: A Romance of Many Dimensions* (London, 1884), an early science fiction work full of Shakespeare references and set in a two-dimensional universe. The significance of *Flatland* to many who work in higher-dimensional geometry is shown by a recent scholarly edition sponsored by the Mathematical Association of America (Cambridge, 2010 – editors William F. Lindgren and Thomas F. Banchoff), and its use in physicist Lisa Randall's account of theories of multiple dimensionality, *Warped Passageways* (New York, 2005), pages 11-28.

[17] Most standard statistics textbooks give accounts of 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.

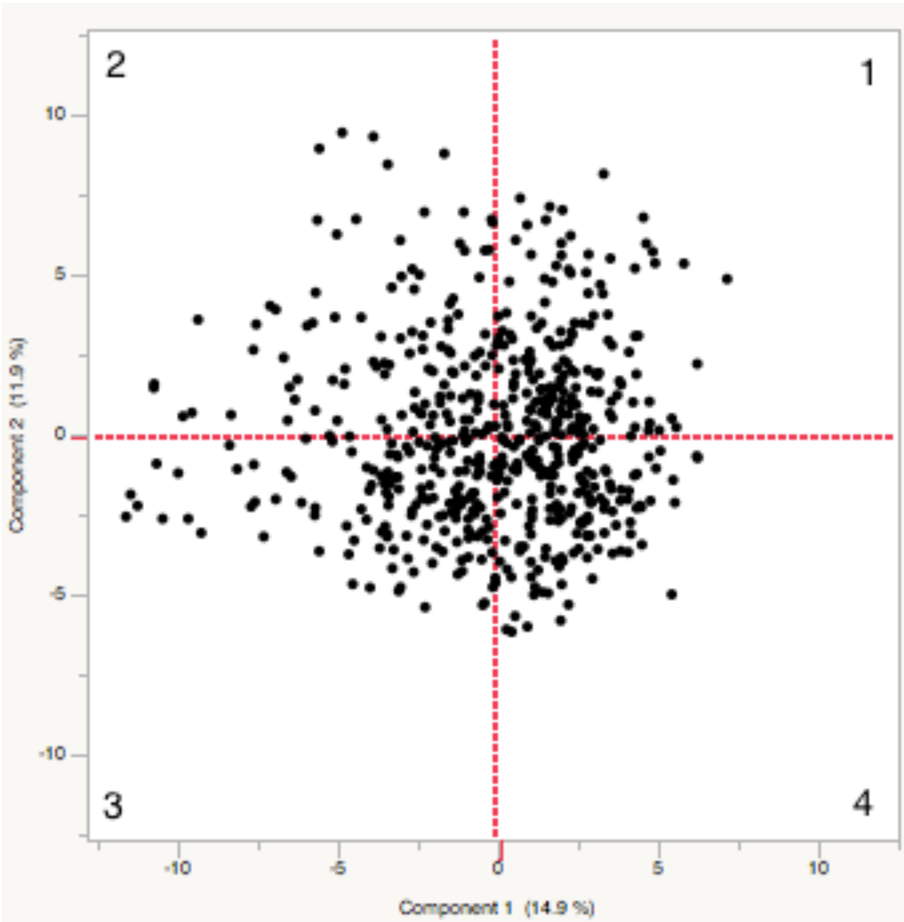
dimensions as possible: of course, they actually run at various angles to existing lines, but the smaller the angle, the more of the information on the original line they retain. Mick Alt describes a principal component as like the central handle of an umbrella whose spokes (the multiple dimensions) have been bashed about by the wind: some of them are bent away from the central pole, at large angles; some are still relatively tightly folded up. The central pole points in roughly the same direction as the tightly folded spokes, so it 'retains' a lot of the information they do – it is a relatively good guide to the directions they point in - while it is a relatively bad guide to where the broken spokes are pointing. Imagine multiple sets of spokes pointing in multiple dimensions, and multiple 'handles' trying to summarize the directions, and you have something like PCA.

PCA software works by first calculating the 'handle' that summarizes the largest possible number of spokes. This is termed 'Principal Component 1' (PC1). The software then moves on to calculate the handle that summarizes the largest amount of the remaining data (spokes). This is termed PC2. These first two PCs should capture a lot of the total variation: especially if the data has relatively straightforward patterns, so that many of the spokes point in roughly the same directions (for example, if the presence or absence of one set of things is a reliable predictor for the presence or absence of another set of things). Items in the data set (plays) have scores on each PC, just as we began by plotting plays onto a single line of values for a LAT. So it is possible to use the PCs as axes for a two-dimensional graph, and plot plays in 'PCA space': a two-dimensional summary of a much more complex space.

So, we build-up a hyper-dimensional space containing 591 plays, and then collapse it into two dimensions. Something we can't imagine, becomes something we can. There is a loss of course:

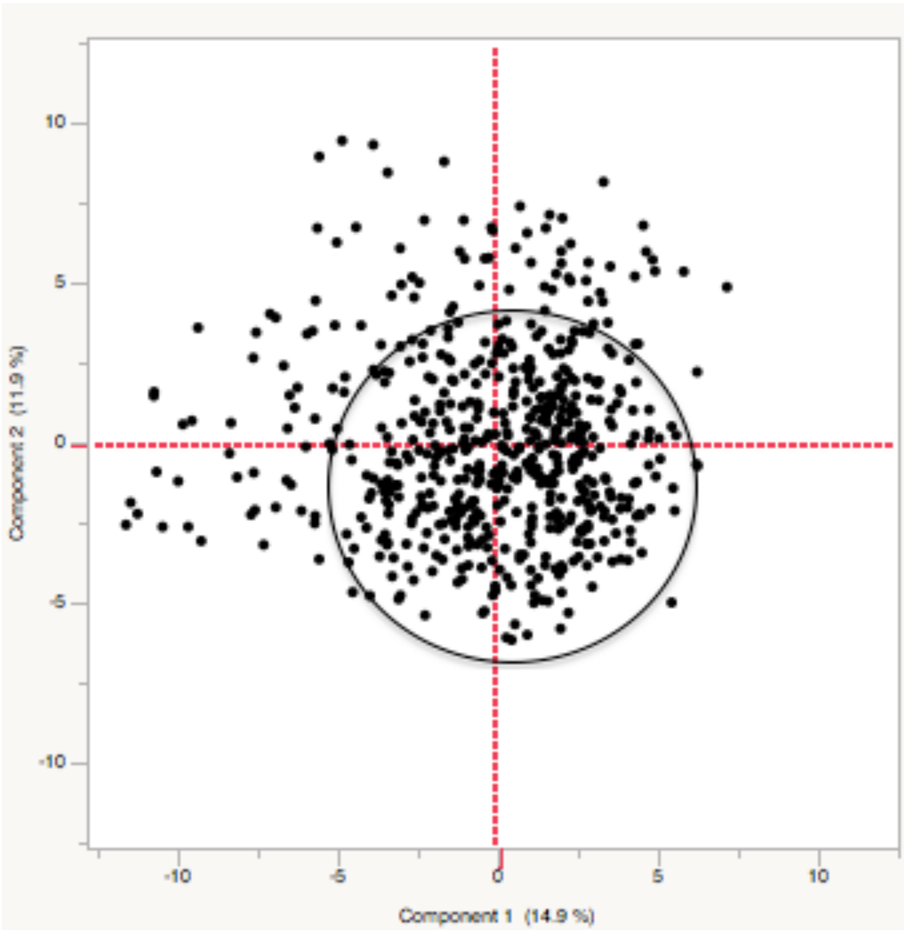
reducing 74 dimensions to two takes away a huge amount of information. Imagine looking at the night sky through a telescope: the milky way appears as a thick band of crowded stars, some apparently right next to each other, or on top of each other. But in many cases, these stars are actually huge distances apart: if we could fly up in three dimensional space and see the galaxy from above, we would see this distance. Because we look at what is essentially a two-dimensional representation of space, we lose this aspect of the data. We need to remember when looking at plots of PCA space that something similar has happened: we are looking at one way of slicing through the complex data-space – a way designed to give us as much information as possible, but one which necessarily misrepresents (if only by omission). Figure 9 shows us all 591 plays in our Early Modern drama corpus plotted against the first two principal components identified by our statistical software. The plays appear as black dots, and their position is fixed according to their scores on PC1 (the horizontal axis, which accounts for 14.9% of the total variation in the data), and PC2 (the vertical axis, which accounts for a further 11.9%: so we are getting at 26.8% of the possible information here). Note that the axes of the graph meet in the centre, where both have a value of 0, dividing PCA space into four quadrants (numbered anti-clockwise from top right, which is quadrant 1: to top left, quadrant 2; bottom left, quadrant 3; and bottom right, quadrant 4).

Figure 9: All 591 plays plotted in PCA space, with quadrants numbered



In terms of reading this image, note that most plays fall in a fairly tight circle which could be drawn with its centre slightly displaced diagonally down and right from 0,0 into quadrant 4. This is especially true of quadrant 4, where almost no plays would fall outside this circle. Quadrant 2, on the other hand, has a relatively looser distribution, with quite a few plays outside the circle, and a much less dense patterning of plays in the area of the circle that falls inside the quadrant. Quadrants 1 and 3 have similar distributions to each other: dense distributions of plays in the circle, and looser distributions outside it on the areas bordering quadrant 2. Overall, we can say that there is a central core of densely distributed plays, and a looser cloud of outlier plays in the areas surrounding it, especially in quadrant 2, but spreading out into quadrants 1 and 3.

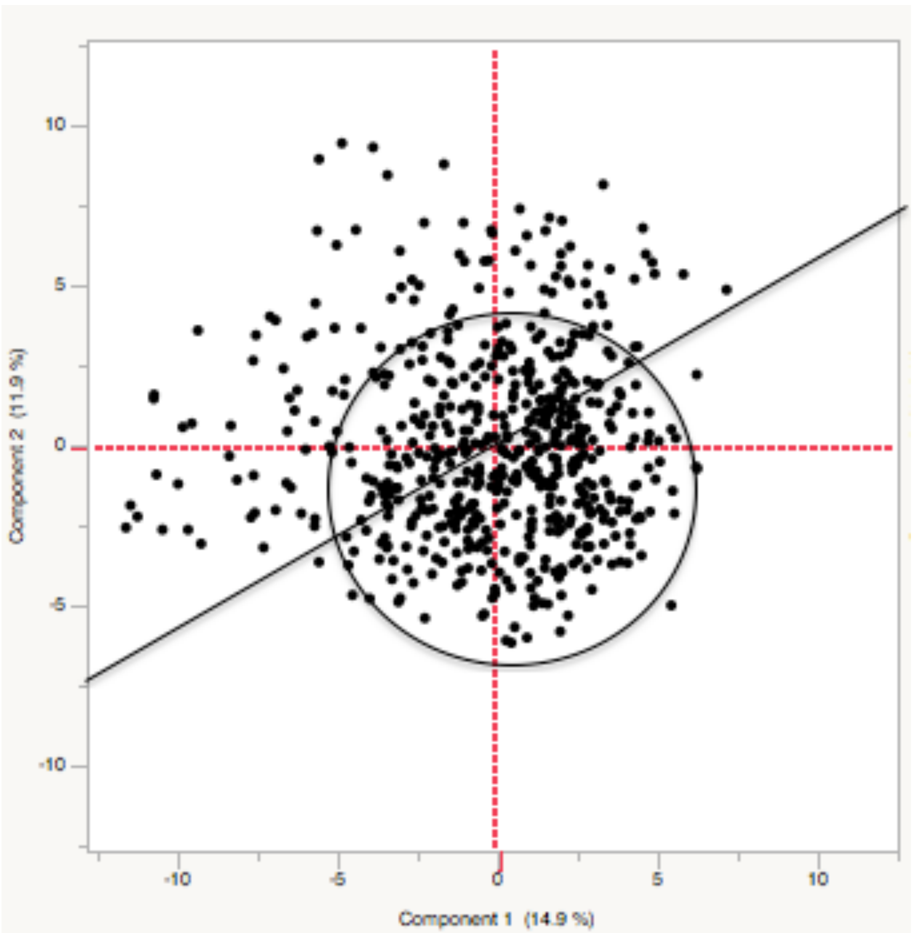
Figure 10: All 591 plays plotted in PCA space, with suggested 'core' circle



What does this tell us? The numbers lying behind this graph record the frequency of 74 linguistic features in each of the 591 plays. Proximity in the PCA space represented in the graph therefore equates to linguistic similarity – and distance equates to linguistic difference. Plays close to each other in the graph (in the same quadrant, for example) will have similar frequencies of the linguistic features. Plays a long way from each other (diagonally opposite each other across two quadrants, for example) will have very different frequencies.

So one conclusion from the graph is that while some Early Modern plays differ a lot from each other (the extreme plays in quadrant 2 versus the one obvious outlier in quadrant 4 for example), most fall into a central, 'core', area of the graph, suggesting a good degree of linguistic consistency. Furthermore, we can say that this central area is continuously populated: there are no obvious groups, or blank areas. Given the available linguistic space of this core circle, Early Modern writers occupy all of it. Beyond the core, variation occurs only in certain areas of linguistic space: quadrant 2, and the bordering areas in quadrants 1 and 3. If we drew a diagonal line across the graph from bottom left to the top right, almost none of the outlier plays would be on the right-hand side of the line (see Figure 11).

Figure 11: All 591 plays plotted in PCA space, with suggested diagonal line showing the areas in which linguistic variation is greater (upper left of the graph)

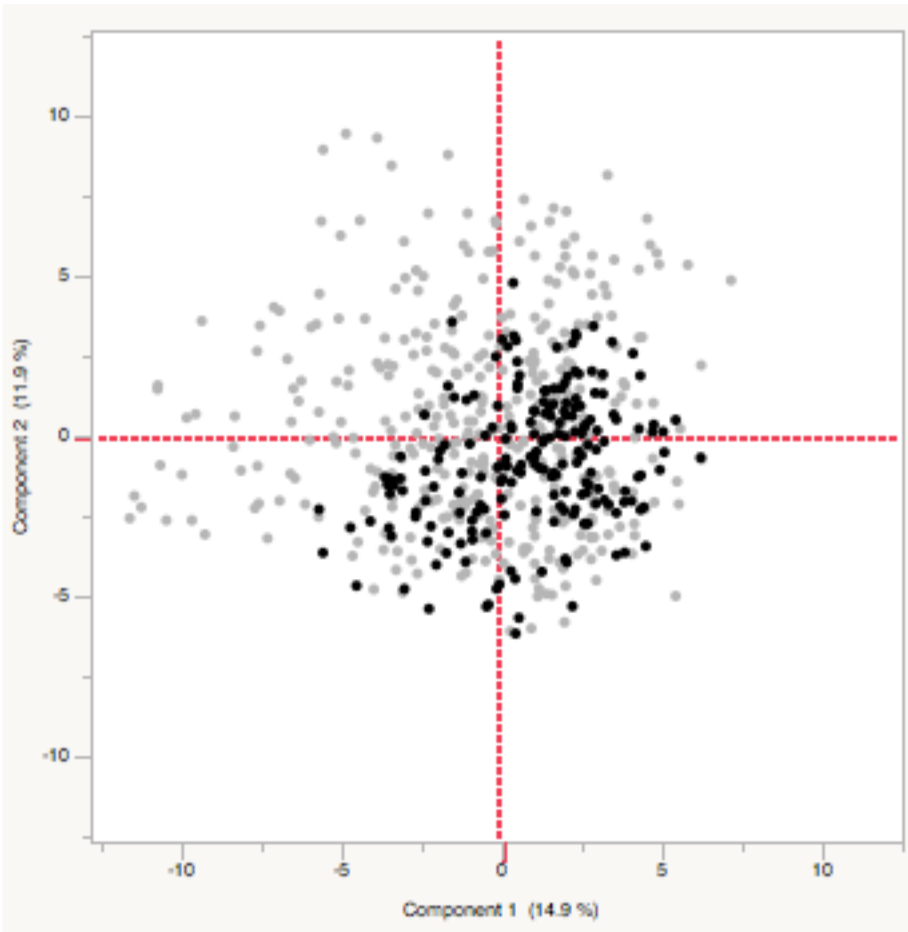


We began this section of our chapter with a focus on the three major companies identified in Figure 3, and our research question, 'Do Theatre Companies, or any other groupings of writers, have recognizable linguistic styles?'. Having established our methodology (tagging with Docuscope, and analysis using PCA), we can now address this directly. A series of initial tests were performed on the plays associated with the three major companies, defining each by the plays written by the main writers associated with each company (The King's Men: Fletcher, Shakespeare, Massinger; Henrietta's: Heywood, Shirley, Brome; and Prince Charles' Men: Dekker, Ford, Rowley, Webster). However, we could find no clear linguistic differences between the major companies.[18]

What we did find, however, was something perhaps more interesting than the existence of 'house styles'. During the testing procedure, we amalgamated all of the plays from the three major companies: a corpus of 211 plays from the total of 591. When we looked at the positions of this smaller corpus in PCA space, we saw something we would not have predicted. Figure 12 repeats Figure 9, this time with just the 211 major company plays in black, and the other plays in grey, showing the location of all the major company plays in PCA space.

[18] There are some minor differences between the major companies, and we would like to do more work on the minor ones, especially the Children's companies.

Figure 12: The 211 plays by major writers associated with the three major companies (black) plotted in PCA space against the remainder of the 591 play corpus (grey)



This figure is striking: the 211 major company plays are almost entirely in the area of the core circle; none is obviously in the outlier cloud in the upper left half of the graphical space. Moreover, the 211 plays seem to avoid quadrant 2, the area of greatest looseness in the drama corpus as a whole, and the quadrant with the smallest area of core circle in it. It is important to remember that we selected the 211 plays in this group purely on the basis that their author had to be an affiliated author to one of the three major theatre companies, with a large body of plays: we did not select on the basis of linguistic behaviour. A reasonable expectation of the likely distribution of the 211 plays selected would have been that they would pattern evenly across the whole distribution of the complete 591 play corpus, with an appropriate number in quadrant 2, and in the outlier cloud. There was no reason to suppose that the playwrights selected would show such tightly grouped linguistic behavior: indeed, given the wide differences in date in the group, we might have expected a very wide distribution.

We draw two preliminary conclusions from this data visualization:

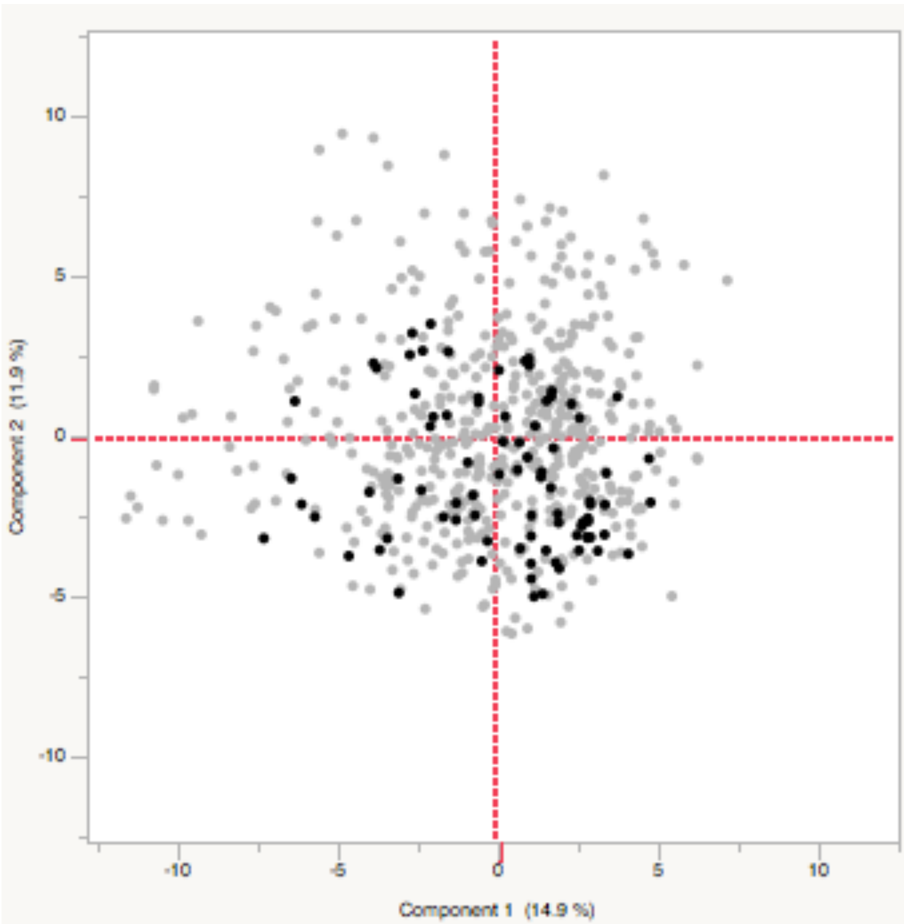
(1) compared to the corpus of Early Modern dramatists as a whole, dramatists writing for the major companies have a consistent and relatively constrained style: a 'target' they aim for and hit with regularity

(2) within the 'target style' one quadrant of PCA space (2) is dis-favoured, with relatively few major company plays appearing there

So, although we can find no clear evidence for communities of practice based on and distinguishing individual theatre companies, this finding suggests that there is a community of practice amongst all of the professional writers attached to the three major theatre companies: the language choices of such writers consistently fall within a well-defined 'target' of linguistic practice. We cannot claim, on this evidence, to have isolated two styles, one used by major company playwrights, or 'core' writers, opposed to one used by minor, non-core writers. Many of the plays by non-core writers also occur in the central circle (they are there, but greyed out, in Figure 12). We can claim, however, that core playwrights attached to the three major theatre companies form a community of relatively coherent, restricted, practice: their plays vary across quadrants 1, 3, and 4, but they vary within certain limits. A hypothesis is that part of the process of becoming a successful playwright is learning to hit this linguistic target.

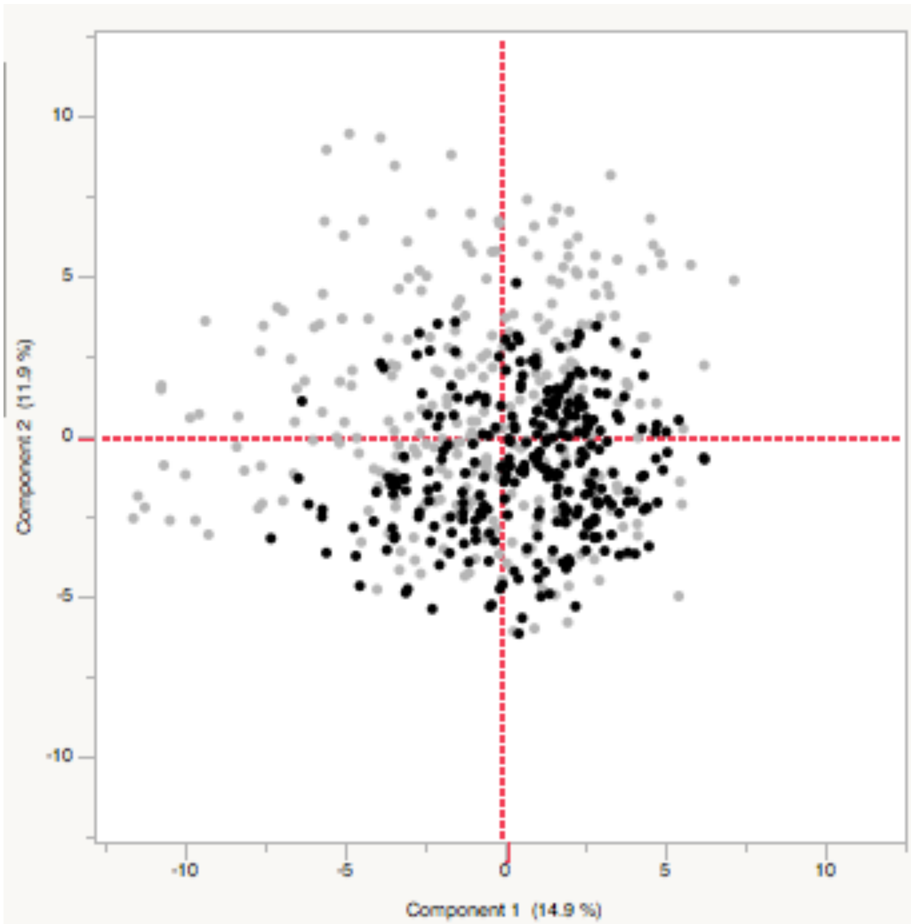
Before we make too much of this, we should remember that our definition of 'core' group excludes several very significant professional playwrights: Jonson, Lyly, Marlowe, Middleton, Chapman, and Marston, for example. Intuitively, there are good reasons for wondering if one or more of these sometimes stylistically striking writers might turn up in the 'outlier' zone to the left of the diagonal line in Figure 11. We have tested for this by adding their plays to the analysis. In the following Figure 13, 86 plays by these writers are represented by black dots, with the rest of the corpus (including those of the 'core' group) grey.

Figure 13: The 86 plays by Jonson, Lyly, Marlowe, Middleton, Chapman, and Marston (black) plotted in PCA space against the remainder of the 591 play corpus (grey)



We would suggest that this is another striking result: we have added 86 plays to our analysis, again by major playwrights, but including a significant date and stylistic range (Lyly to Marston; Marlowe to Middleton). Still the majority of plays fall within the target circle at the centre of the graph. Certainly, we now have more plays in quadrant 2, and some outlier plays in quadrant 3 (often by Marlowe, perhaps not surprisingly), but overall the pattern remains the same: plays by established professional playwrights have a consistent, and constrained, linguistic style. It certainly looks as though major playwrights have a stylistic community: not an exclusive one, since many plays by non-major writers fit in the circle; but a proscriptive one: write outside these bounds, and you will not be employed for long in the Early Modern theatre. As a summary, Figure 14 shows all 297 plays by major playwrights in black, with other plays in grey.

Figure 14: The 297 plays by major playwrights (black) plotted in PCA space against the remainder of the 591 play corpus (grey)



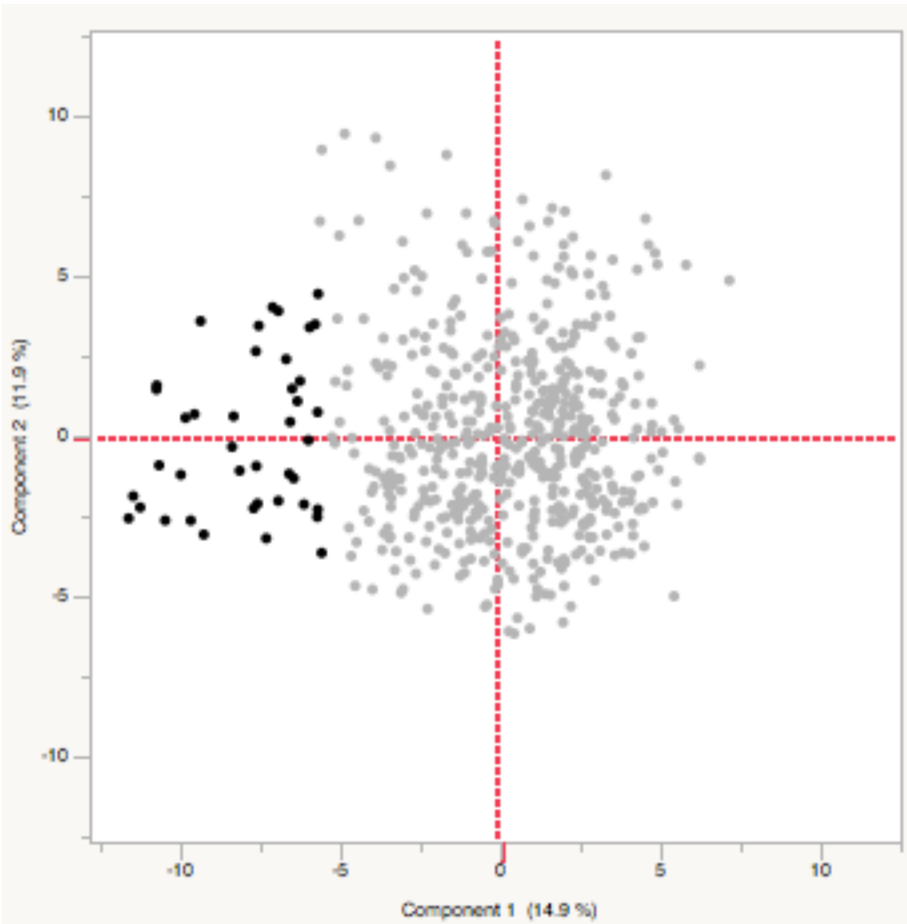
This analysis raises two questions:

(1) How does this graph relate to the actual linguistic choices writers make?

(2) What is unusual about the plays in the ‘outlier cloud’ beyond the inner core?

As a way of answering both of these questions, let’s look at the outlier plays on the left side of the graph, lying across quadrants 2 and 3. These are highlighted in Figure 15.

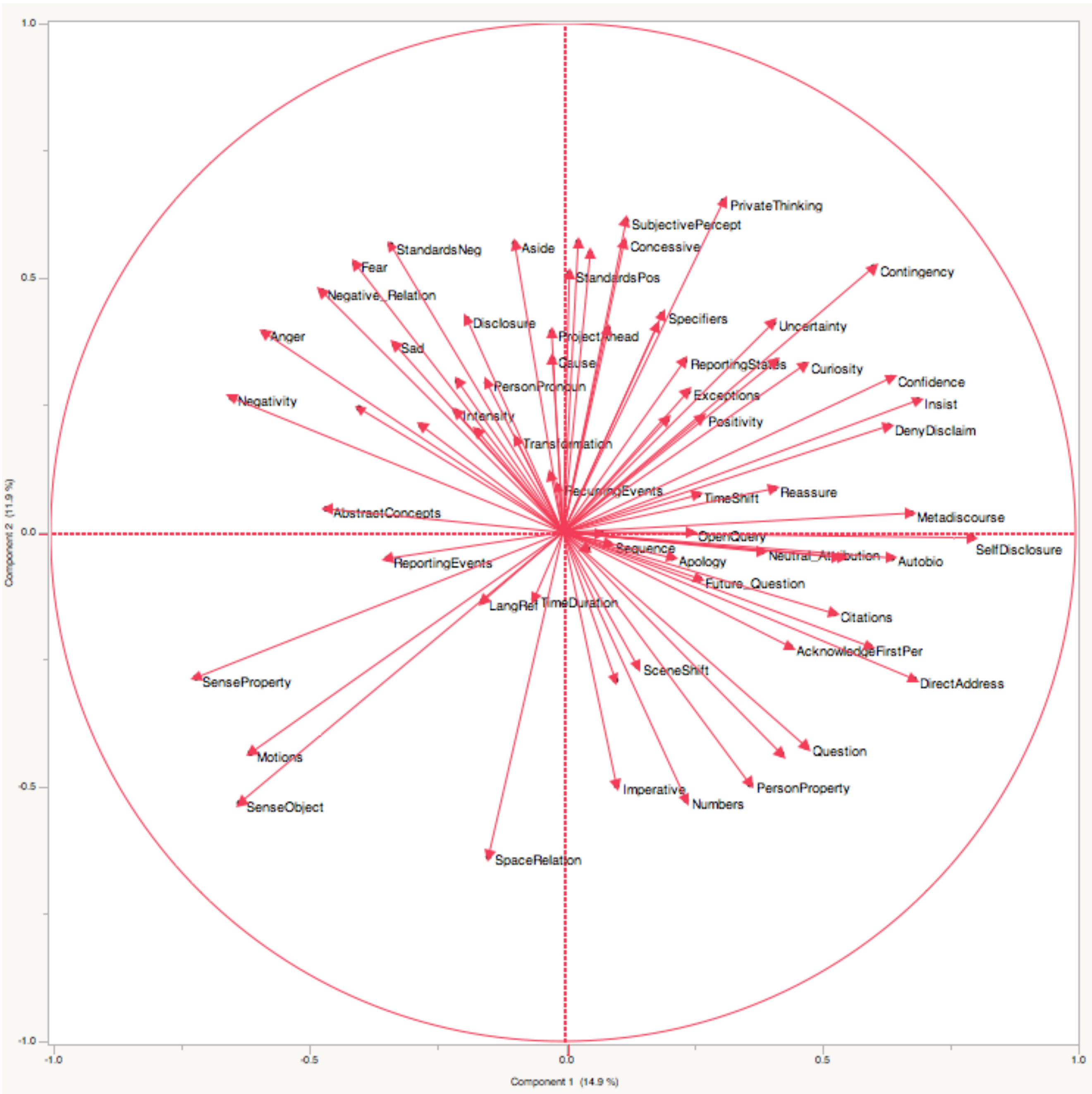
Figure 15: The left-outlier plays (black) plotted in PCA space against the remainder of the 591 play corpus (grey)



These plays have been pulled into the left side of the graph because they score lower on PC1 (plotted on the horizontal axis) than core plays. Conversely, in terms of their scores on PC2 (the vertical axis), they score the same as core plays (they lie within the range from +5 to -5). So the variation between these plays and the core group is being captured on PC1. What are these plays doing to make their scores diverge from the core group in this direction?

Individual plays are positioned on the map according to their scores (or 'loadings') on PC1 and PC2. These loadings summarise the relationships between the raw frequency counts for 74 variables (LATs) counted by Docuscope: and we can map those variables onto the same space, giving an indication of what LATs plays use, and avoid. Figure 16 shows a projection of the LATs into the same PCA space: each LAT is represented by an arrow, with the length of the arrow indicating the strength of the relationship between the LAT and the PC. Plays plotted in a particular quadrant of the graph are there because they use more of the LATs which are projected into that quadrant – and less of the LATs in the opposite quadrant (because of the way the maths works, we can ignore the positive or negative values on the axes: length in any direction indicates high frequency of use).

Figure 16: The 74 Docuscope LATs used in this study with their projections into PCA space



The left-lying outlier plays are positioned there because they use relatively more of the LATs that project furthest into the left hand space of the graph – and there are two sets of LATs pulling these plays into this space, one set in quadrant 3; one set in quadrant 2.

In quadrant 3, the LATs projecting furthest capture language associated with denoting, describing, and locating objects in space: 'Sense Object', which tags nouns; 'Sense property', which tags adjectives describing physical characteristics; and 'Motions', which picks up language encoding movement. In quadrant 2, the LATs projecting furthest identify a set of linguistic features coding for negative and emotional language: 'Standards Negative', which tags negatively judged terms and concepts (for example, 'disease', 'unworthy', 'disorder'); 'Fear' which tags words associated with fear, such as 'fear', 'threatening', 'terror'; 'Negative relation', which tags words implying confrontation ('you cannot', 'your tears', 'don't' you'); 'Anger', which tags words associated with or referencing anger ('angry', 'vengeance', 'slaughter', 'coward'); 'Negativity', which tags negative emotion ('gloom', 'distrust', 'abhor' 'villain'); 'Sad', which tags words expressing sadness ('cry', 'pity', 'despair', 'no hope'). To give examples of these in context, we'll take an extract from the play that has the most extreme value on PC1 – the play lying furthest left in the graph. This play turns out to be a translation of Seneca's *Hippolytus* by John Studley.

The following speech by Phaedra gives a flavor of the language of the play:[19]

[19] The text as printed here comes from the HTML file generated by Visualizing English Print's in-development web-based version of Docuscope. The HTML file is derived from the EEBO-TCP file of the text: as readers will see, the EEBO-TCP files contain numerous gaps – marked <ENCODE_ERROR> here. The EEBO-TCP files will be improved as more people work on them – for example, the F21 project based at the Folger Shakespeare library proposes to clean up the drama texts –

I know the truth ye teach
Oh Nurse, but fury forceth me at worse things to reach:
My mind even wittingly to vice falls forward prove and bent
To bolesome counsel back again in vain it does relente:
As when the Norman tugges and toils to bring the freighted Bark
Against the striving stream, in vain he loses all his cark
And down the shallow stream perforce the Ship does headlong yield,
Where reason preaseth forth, there fighting fury wins the field,
And bears the swinging sway, and crank Cupidoes puissant might
Tryumpheth over all my breast this flighty <ENCODE_ERROR> wight
And puissant potestate throughout the world does hear the stroke,
And with unquenched flames does force <ENCODE_ERROR> kindled breast to smoke,
The Battelbeaten Mars hath felt these bitter burning brands,
And eke the God hath tasted these whose fervent fiery hands,
The thumping thunder bouncing bolts three forked wise does frame,
And he that ever busted is about the furious flame,
In smoltring Furnace raging hot on dusky top so <ENCODE_ERROR>

---

but any large data set has a 'dirty' quality. It is counter-intuitive for literary scholars, who are often, rightly, concerned with fixing minute errors in texts, but statistical work on the scale we are working at here does not need perfect data: as long as errors are more or less random throughout the data, then the results will be robust. Aside from the gaps, there are some obvious errors in transcription (the original text is set in black letter, which the TCP transcribers, like many of us, find hard to read) – 'greic vous' in the fifth last line should be 'grievous', for example – but there is not time to re-edit all the texts to remove these errors – and for our purposes, there is no need. The text is good enough.

Of foggy <ENCODE_ERROR> mount: and with such slender heat does fry,
And Phoebe himself that weldes his dart upon his twanging string,
With aimed shaft directly driven the wimpled Lad does sting.
With pour he scours along the Earth and Marble Sky awayne.
Lust favouring folly filthtly did falsely forge and <ENCODE_ERROR>
<ENCODE_ERROR> for a God: and that he might his freedom <ENCODE_ERROR>
Ascribes the name of feigned God to shittel bed lame rage.
Erycina about the world does send her roving page,
Who gliding through the Azure skies with slender jointed arm
His perilous weapons wields at will, and working griec vous harm.
Of bones and stature being least great might he does display
Upon the Gods, compelling them to crouch and him obey.
Some Brainsick head did attribute these things unto himself,
And Venus Godhead with the bow of Cupid little self.

(John Studley (tr.) *Hippolytus* from: *Seneca his tenne tragedies, translated into Englysh* 1581, London; act 1, scene 2; original book pagination pp. 58-59; JISC Historic Books (page image number 63), http://www.jischistoricbooks.ac.uk/Search/?bibnumber=STC%20(2nd%20ed.)%2022221.&spage=63 (accessed 7.2.2014))

The lumbering fourteeners and insistent alliteration mark this out stylistically as of its time: no one used to reading the lithe iambic lines perfected by Shakespeare and Marlowe in the 1590s will be surprised that this play is plotted at such a distance from the core of Early Modern theatre. But the data visualized in Figures 9-15 contains nothing about metrics or alliteration: Docuscope does not count syllables or repeated first letters.[20] The fact that Docuscope can pick these plays out tells us that there are more differences than the purely formal ones of metrics and sound patterning marking them as old-fashioned. As well as those formal features, the play's language departs from the core style of Early Modern drama in its increased use of two sets of semantic features counted by Docuscope: the spatial and object-centred features tracked in quadrant three, and the negative, emotional ones tracked in quadrant two.

To give you a sense of the density of tagging, here is the passage again, with *all* of the words and phrases tagged by Docuscope in the LATs identified as raised in this style in bold, and then collected together by LAT at the end of the extract:

> I know the truth ye teach
>
> Oh Nurse, but **fury** forceth me at **worse** things to reach:
>
> My mind even wittingly to vice falls forward prove and bent
>
> To bolesome counsel back again **in vain** it does relente:
>
> As when the Norman tugges and **toils** to bring the freighted **Bark**
>
> Against the striving **stream, in vain** he loses all his **cark**
>
> And **down the shallow stream** perforce **the Ship** does **headlong** yield,
>
> **Where** reason preaseth **forth**, there fighting **fury** wins the **field**,

---

[20] Visualizing English Print's web implementation of Docuscope will allow users to upload their own rule-sets, so it will be possible to adapt Docuscope's tagging rules, or replace them completely.

And **bears** the **swinging sway**, and crank Cupidoes puissant might

Tryumpheth over all my breast this flighty <ENCODE_ERROR> wight

And puissant potestate **throughout the world** does **hear the stroke**,

And with unquenched **flames** does force <ENCODE_ERROR> **kindled** breast to **smoke,**

The Battelbeaten **Mars** hath felt these **bitter burning** brands,

And **eke** the God hath **tasted** these whose fervent **fiery hands**,

The **thumping thunder bouncing bolts** three **forked** wise does **frame**,

And he that ever **busted** is about the **furious flame**,

In smoltring **Furnace raging** hot on **dusky top** so <ENCODE_ERROR>

Of **foggy** <ENCODE_ERROR> **mount:** and with such **slender heat** does **fry**,

And **Phoebe** himself that weldes his **dart** upon his **twanging string**,

With aimed **shaft** directly driven the wimpled Lad does **sting**.

With **pour** he scours **along the Earth** and **Marble Sky** awayne.

Lust favouring **folly** filthtly did **falsely** forge and <ENCODE_ERROR>

<ENCODE_ERROR> for a God: and that he might his freedom <ENCODE_ERROR>

Ascribes the name of feigned God to shittel **bed lame rage**.

Erycina about the world does send her **roving page**,

Who **gliding** through the **Azure skies** with **slender jointed** arm

His **perilous** weapons wields at will, and working griec vous **harm**.

Of **bones** and **stature** being least great might he does **display**

Upon the Gods, compelling them to **crouch** and him obey.

Some **Brainsick head** did attribute these things unto himself,

And Venus Godhead with the **bow** of Cupid little self.

Tagged words and phrases sorted by LAT:

Sense Object: field, flames, Mars, thunder, frame, flame, heat, Phoebe, string, shaft, Marble, bed, page, skies, bones, head, bow (raw frequency of this LAT in the play: 674)

Sense Property: shallow, burning, tasted, thumping, forked, dusky, slender, twanging, Azure, slender, stature (raw frequency of this LAT in the play: 335)

Motions: swinging, sway, kindled, bouncing, mount*, dart*, crouch (raw frequency of this LAT in the play: 238)

*Docuscope miss-tags these words: in context, they are nouns (‘mount’ = mountain), but it assumes they are verbs (because it is a simple string-matching tagger without grammatical information), hence their association with movement.

Negativity: the stroke, bitter, eke, busted, raging, sting, lame (raw frequency of this LAT in the play: 358)

Standards Negative: ERROR*, folly, falsely (raw frequency of this LAT in the play: 249)

Anger: fury, furious, rage (raw frequency of this LAT in the play: 50)

*DS tags the transcriber’s encode error note: these are likely to be more frequent in black letter texts, so may skew the results. Public implementation of DS will fix this: probably that errors are sufficiently repeated in all texts so that is not affecting results significantly.

Just looking at the bold text here gives us a glimpse of a new way of doing literary study. We are not basing our argument on a few, cherry-picked, examples. Our argument is based on a very large number of instances, no one of which is more important than the rest. We could easily have chosen another section of the play to illustrate the LATs – and we could easily lose any one of the examples given here without the loss affecting our claims about the language of the play. This is why we can speak of an obviously imperfect transcription as being 'good enough': our evidence is the very high frequency of a set of items – one or two, or even ten or twenty either way, would make little difference to the location of the play on the graph, or our conclusions. This is a strange world for literary scholars, but one we should get used to: and it is not, pace Franco Moretti, a new world of distant reading which banishes the old world of closely read phrases and passages.[21] The two approaches can work together and inform each other: 'The **thunder thumping bouncing bolts** three-**forked** wise does **frame**' is not only a tag-rich line contributing to the statistics, but one we are tempted to savour as an example of what fourteeners are perhaps best at – rugged, robust language which conveys vividly the vigour of the natural world. But thanks to statistics and distant reading, we can savour this line in the context of knowing that its focus on objects ('thunder'), their properties ('thumping', 'forked'), and motions ('bouncing') is part of a linguistic strategy common to a range of texts written around the same time.

So we can say that the outlier plays are pulled to the left of the graph because they use higher frequencies of about six LATs, as described above. Although we can pull out examples of these LATs to show the kind of linguistic effects they create, our argument here is not based on individual cases, but on the totality of cases: the general trend picked up by the statistics. This is an unusual, though not entirely unknown, approach to literary analysis.[22] However, at the same time as the outlier plays are

[21] Franco Moretti, *Distant Reading* (London, 2013).
[22] For a discussion of quantification in literary analysis, see Jonathan Hope and Michael Witmore, 'The Language of *Macbeth*' in Ann Thompson (ed.), *Macbeth: The State of Play* (London: 2014), pp. 183-208.

pulled to the left because of the linguistic choices their authors make in using certain features at a high rate, they are pushed away from the right hand side of the graph because of a further set of linguistic choices: this time to avoid particular features. The plays are pushed away from the right side of the graph because they use low frequencies of LATs that extend strongly there. Figure 16 shows that these LATs are 'Contingency', 'Confidence', 'Insist', 'Deny Disclaim', 'Self Disclosure', 'Autobiography', 'First Person', and 'Direct Address'.

Before we discuss these LATs, let's reflect on what we are able to do here thanks to digital tools: electronic tagging, searching, and data analysis allow us not only to identify what is *in* the texts we are studying, but also what is *absent*. Literary arguments have certainly been constructed on the basis of what texts lack in the past, but not in such a systematic way – and not with such confidence in terms of placing texts within their period and generic context. As literary scholars become more used to dealing with electronic corpora, this aspect of study is likely to become more and more important: arguably, genres and groups of texts are constructed by what they *do not* do even more than by what they *do* do.

So what is it that these outlier texts are avoiding? As with the LATs that the outlier plays use with increased frequency, we can group the reduced LATs into two broad categories. The first group, projecting into quadrant 1, capture language that sets up a speaker's relationship to what is being spoken. For example, 'Contingency' tags words which express possibility: 'may', 'might', 'had I', 'chance', 'could' – so rather than making declarative statements of fact about the world ('He did it'), speakers are making clear that their claims are contingent ('He may/might/could have done it', 'There is a chance he did it'). 'Confidence' similarly tracks speaker expressions of the degree to which they have confidence in what they say: 'know', 'assure', 'absolute', 'clearly'. Paradoxically, the assertion of confidence has the effect of underlining the contingent nature of knowledge: compare, ‘This is true’ with ‘I assure you, I know that this is absolutely true’. ‘Insist’ captures modal and other verbs used to express confidence, so can be seen as functioning in the same way as ‘Confidence’: ‘must’, ‘should’, ‘insist’. For example: ‘I insist that this must be true’. ‘Deny Disclaim’ and ‘Meta-discourse’ track language which marks systematic argumentation: in the case of ‘Deny Disclaim’, as the name of the LAT indicates, the language contradicts previous assertions (‘neither’, ‘nothing’, ‘no’, ‘never’, ‘cannot’: ‘No, that cannot be true’); in the case of ‘Meta-discourse’, the language guides us through an argument: ‘too’, ‘we shall’, ‘but there is’, ‘further’, ‘moreover’, ‘aforesaid’). So these LATs are concerned with qualifying statements about the world in terms of the speaker’s knowledge, and contradicting other speakers’ claims. We should note that they can thus be opposed to the concrete, world-based language of ‘Sense Object’ and ‘Sense Property’ of quadrant 3.

The second group of under-used LATs project into quadrant 4. They are concerned with tracking language which marks rapid oral interchange and self-revelation. The LATs which capture self-revelation are ‘Self Disclosure’, which tags first person forms appearing with a particular set of verbs (‘I am’, ‘I think’, ‘I feel’, ‘I believe’, ‘I confess’) and ‘Autobiography’, which searches for first person forms either attached to verbs or other words with an implication of recollection (‘I have been’, ‘I was’, ‘when I’) or attached to nouns which indicate a relationship (‘my daughter’). The association of these LATs with those of quadrant 1 is clear: here too we are concerned with language making explicit the presence of the speaker, and the relation of them to what is being said (broadly speaking, the truth value of what is said is being made explicit in some way: instead of ‘I know this’, we have ‘I know this because…’).

Accompanying these LATs of self-revelation, we have a set of LATs which capture language used to represent speech: 'Direct Address' tags second person pronouns ('you', 'you are', 'prithee', 'thy', 'thou') and thus tracks dialogue exchanges; 'Question' tags question words, and punctuation ('what', 'hast', 'is it', '?') – which again are frequent in dialogue involving rapid interchange. 'Oral Cues' (the unlabeled arrow in Figure 16 just beside 'Question') captures words such as 'well', 'my word', 'oh', 'yes', 'ye', 'good morrow', 'ah', 'nay', which are characteristic of writing which attempts to mimic informal speech.

So we can say that the outlier plays avoid a set of strategies which we find consistently in the core group plays: language which focuses on the beliefs and psychological attitude of the speaker to what is asserted, language which reveals the past of the speaker, and language that mimics the rapid interplay found in actual speech situations. What is our literary interpretation of this result? We would suggest that the analysis reveals that the outlier plays share a heightened language, often using vivid adjectives ('Sense Property': shallow, burning, tasted, thumping, forked, dusky, slender, twanging, Azure), and language which encodes strong emotion ('Negativity': the stroke, bitter, eke, busted, raging, sting, lame; 'Standards Negative': folly, falsely; 'Anger': fury, furious, rage). Even 'Motions', designed to detect language recording movements in space, is emotionally charged, because of the figurative use of movement to express strong feeling. It is tempting to see these outlier plays as occupying a space of over-writing, where inexperienced writers are straining for emotional effects, and over-using language which carries an emotional charge. Better writers, including all who make a living as professional playwrights in the period, know that genuinely effective writing omits and suggests: their use of these charged LATs is therefore reduced in comparison, and their plays are concentrated in the core area of the graph. Conversely, the core writers are more focused on the psychology of their characters, and representing lively spoken interchanges.

What we have here is an initial investigation of a data set which, Shakespeare and a few others aside, is surprisingly little known. The data itself, as we have seen, is imperfectly transcribed. The metadata we have is also imperfect. But even now, we can begin to identify broad patterns in the data we were not aware of before. We can begin to ask questions about similarity and difference in the corpus, and search for communities, in ways that were impossible before electronic text analysis and data visualization. The techniques of tagging and data visualization and analysis are relatively well-understood in corpus linguistics and statistics. But those fields can tell us nothing about the right questions to be asking this data: literary scholars are the only people who can do that. And literary scholars are the only people who can begin to interpret the results of visualization of this data. This is good news for the future of literary scholarship; though we are going to have to learn some statistics.