Actes des congrès de la Société française Shakespeare

31 | 2014
La langue de Shakespeare

Quantification and the language of later Shakespeare

Jonathan Hope and Michael Witmore

Christophe Hausermann (ed.)

Electronic version
URL: http://journals.openedition.org/shakespeare/2830
DOI: 10.4000/shakespeare.2830
ISSN: 2271-6424

Publisher
Société Française Shakespeare

Printed version
Date of publication: 1 May 2014
Number of pages: 123-149

Electronic reference
Jonathan Hope and Michael Witmore, « Quantification and the language of later Shakespeare », *Actes des congrès de la Société française Shakespeare* [Online], 31 | 2014, Online since 01 May 2014, connection on 10 December 2020. URL : http://journals.openedition.org/shakespeare/2830 ; DOI : https://doi.org/10.4000/shakespeare.2830

QUANTIFICATION AND THE LANGUAGE OF LATER SHAKESPEARE

Jonathan HOPE and Michael WITMORE

In this paper we consider the status of quantitative evidence in literary studies, with an example from our own work using the software package Docuscope to investigate chronological 'periods' in Shakespeare's career. We argue that quantitative evidence has a function in literary studies, not as an end in itself, but as a starting point for traditional interpretative literary analysis. In our example, we show that linguistic analysis suggests three periods in Shakespeare's career, defining a 'period' as a group of plays with similar linguistic features. We focus on the latest period, as this is the largest, and suggest that the 'late style' of Shakespeare may begin much earlier than traditionally thought. We analyse the features that the later plays share, and argue that from the late 1590s Shakespeare can be seen to be adopting features which are (a) closer to speech, and (b) indicate a shift from real-world denotation to a focus on communicating the subjectivity of the speaker.

Cette article s'intéresse aux données quantitatives dans les études littéraires et expose le résultat des recherches que nous avons effectuées avec la suite logicielle Docuscope afin d'établir des « périodes » chronologiques dans la carrière de Shakespeare. Nous pensons que les données quantitatives ont un intérêt dans les études littéraires, non pas comme une fin en soi, mais comme un point de départ pour l'analyse littéraire interprétative traditionnelle. Dans l'exemple suivant, nous montrerons que l'analyse linguistique suggère l'existence de trois périodes dans la carrière de Shakespeare, une période étant définie par un groupe de pièces présentant des caractéristiques linguistiques particulières. Nous nous intéresserons en particulier à la dernière période, la plus longue, et nous émettrons l'hypothèse que le style adopté par Shakespeare à la fin de sa carrière commence bien plus tôt que ne l'affirme la tradition. Nous analyserons les traits communs des dernières pièces et montrerons que le Shakespeare de la fin des années 1590 adopte des traits linguistiques qui sont (a) plus proches du discours, et (b) qui indiquent le passage d'une dénotation du monde réel vers l'expression de la subjectivité de l'interlocuteur.

Quantification has a bad name in literary studies, and especially in the study of Shakespeare where, historically, it was associated with the excesses of the 'distintegrators', and the madness and infighting of authorship studies. Recently, however, thanks largely to the work of figures such as Brian Vickers, Mac. P. Jackson, Ward Elliott and Robert Valenza, and Hugh Craig, attribution studies in Early Modern literature have reached a level of methodological respectability and seriousness. This has been paralleled by a similar advance in the methods of attribution studies more generally.[1] There are still debates,

[1] See: Brian Vickers, *Shakespeare, Co-Author: A Historical Study of Five Collaborative Plays,* Oxford, OUP, 2002; Brian Vickers, "Review Essay: Shakespeare and Authorship Studies in the Twenty-First Century", *Shakespeare Quarterly*, 62, Spring 2011, p. 106-142; Mac. P. Jackson, *Defining Shakespeare: 'Pericles' as Test Case*, Oxford, OUP, 2003; Mac. P. Jackson, "Authorship and the Evidence of Stylometrics", in Paul Edmondson and Stanley Wells, eds., *Shakespeare Beyond Doubt: Evidence, Argument, Controversy*, Cambridge, CUP, 2013, p. 100-110; Hugh Craig, "Stylistic Analysis and Authorship Studies", in Susan Schreibman, Ray Siemens, and John Unsworth, eds., *A Companion to the Digital*

but tellingly these debates are about method, not results: scholars are (mostly) no longer lone riders of personal hobby-horse candidates; they are invested in how to do attribution studies reliably, not trying to prove a case they have emotionally committed to before beginning work.

If quantitative methodology can be shown to have improved, there are still humanities scholars who are wary of what they see as misplaced scientism in the importation of scientific methodology to literary studies. This is a much larger topic than simple attribution, also covering the cognitive turn in literary studies, sharply critiqued recently by Deborah Cameron.[2] Cameron gives a strong defence of the particularity of literary studies as a discipline, though even she rejects the notion that arts and science disciplines are polar opposites. Rather she sees them on a continuum, and it is possible to see work in the social sciences which directly addresses the meeting point of quantitative and qualitative work which might be taken as characterising the two approaches.[3]

Humanities, Oxford, Blackwell, 2004, p. 273-88. The work of Ward Elliott and Robert Valenza is archived at:

http://www.claremontmckenna.edu/pages/faculty/welliott/archived.htm

http://www.claremontmckenna.edu/pages/faculty/welliott/select.htm

For surveys of attribution work outside Shakespeare studies, see (in order of publication): David I. Holmes, "Authorship Attribution", *Computers and the Humanities,* 28, April 1994, p. 87-106; Joseph Rudman,"The State of Authorship Attribution Studies: Some Problems and Solutions", *Computers and the Humanities,* 31, April 1998, p. 351-365; Carole E. Chaski, "Empirical Evaluations of Language-based Author Identification Techniques", *Forensic Linguistics*, 81, Spring 2001, p. 1-65; Harold Love, *Attributing Authorship: An Introduction,* Cambridge, CUP, 2002; Patrick Juola, "Authorship Attribution", *Foundations and Trends in Information Retrieval*, 1, March 2008, p. 233-334; Efstathios Stamatatos, "A Survey of Modern Authorship Attribution Methods", *Journal of the American Society for Information Science and Technology*, 60, no. 3, March 2009, p. 538-56; Moshe Koppel, Jonathan Schler, and Shlomo Argamon, "Computational Methods in Authorship Attribution", *Journal of the American Society for Information Science and Technology*, 60, January 2009, p. 9-26; Moshe Koppel, Jonathan Schler, and Shlomo Argamon, "Authorship Attribution in the Wild", *Language Resources and Evaluation*, 45, no. 1, March 2011, p. 83-94; Joseph Rudman, "The State of Non-Traditional Authorship Attribution Studies – 2012: Some Problems and Solutions", *English Studies*, 93, no. 3, May 2012, p. 259-274.

[2] Deborah Cameron, "Evolution, science and the study of literature: A critical response", *Language and Literature*, 20, February 2011, p. 59-72; in response to: Jonathan Gottschall, *Literature, Science and the New Humanities*, New York, Palgrave Macmillan, 2008; Jonathan Gottschall, "Evolutionary literary studies: A foreword to Norbert Francis's review article 'A modest proposal'", *Language and Literature*, 19, August 2010, p. 301-304; and Norbert Francis, "A modest proposal for a reorientation in literary studies", *Language and Literature*, 19, August 2010, p. 305-317.

[3] David Williamson Shaffer and Ronald C. Serlin, "What Good are Statistics that Don't Generalize?", *Educational Researcher*, 33, no. 9, December 2004, p. 14-25.

If we consider quantification seriously for a moment, however, it will appear that it is not inherently alien to literary studies, or Shakespeare. Shakespeare's writing, after all, is overwhelmingly in a quantified metrical form, where syllables are counted, and ordered according to strict numerical rules, and even the apparent exceptions can be predicted and explained mathematically. Although quantitative metrical forms are not universal in human culture, they are very widespread, so we are justified in thinking of quantification as 'natural' to verbal artistic behaviour.[4]

And quantification does have a respectable past in Shakespeare studies: our chronologies of Shakespeare are largely based on metrical counts made in the late nineteenth century. These studies are notable for the first identification of the late plays: a generic category now accepted by mainstream Shakespeareans, many of whom are unaware of the quantitative basis of the category.[5] This identification is a good example of the status of quantitative evidence in literary studies. The statistical similarities in terms of metrical behaviour that these plays share are undeniable mathematically (they are 'significant' in the technical statistical sense); but this quantitative result only has 'significance' in its broader, non-technical sense if the grouping of plays that emerges from it makes sense to literary scholars. The history of scholarship in the twentieth century shows clearly that this grouping does indeed make very good sense to literary scholars. Paradoxically then, the ultimate test of quantitative methods and results in literary studies is not quantitative, but qualitative: do the numbers give us

---

4 For a linguistic account of Shakespeare's metrics, see Nigel Fabb, *Linguistics and Literature: Language in the Verbal Arts of the World*, Oxford, Blackwell, 1997, p. 37-48 and 51-55. On metrical from as part of verbal behaviour, see Nigel Fabb and Morris Halle, 2012, "Counting in verbal art", in Isabel Jaén and Julien Jacques Simon, eds., *Cognitive Literary Studies: Current Themes and New Directions*, Austin, University of Texas Press, 2012, p. 163-182.

5 These metrical studies were pursued by Frederick Furnivall (a maths graduate before he became a philologist) and the New Shakespeare Society. See, for example, Frederick James Furnivall,*The Succession of Shakspere's works and the use of metrical tests in settling it,* Smith, Elder & Co, London, 1874. For more detail on the recognition that the late plays constituted a generic group, see 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, footnote 2, p. 133-134. For a full account of the critical history of the late plays, see Barbara A. Mowat, "'What's in a name?' Tragicomedy, Romance, or Late Comedy", in Richard Dutton and Jean E. Howard, eds., *A Companion to Shakespeare's Works: The Poems, Problem Comedies, Late Plays*, Oxford, Blackwell, 2006, vol. 4, p. 129-149.

something we can work with? Do they allow new insight or prompt interesting questions?

In this study we are not quantifying metrical features, but linguistic-rhetorical ones. To do this, we use Docuscope, a text-analysis program developed at Carnegie-Mellon University.[6] Originally intended for use in undergraduate writing classes as a means to identify and teach the rhetorical behaviours associated with certain types of writing (instructional, biographical and so on), the program has enabled us to identify the linguistic fingerprints of Shakespeare's genres.[7] Docuscope is essentially a set of dictionaries. It searches a text for strings of characters (letters and punctuation) which it matches to those in its memory. Docuscope strings can include single words:

here
master
cheer
I

phrases:

speak to
the master
take in the
I have
use your

6 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.

7 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]; 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, p. 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, p. 357-390.

words and punctuation:

: what
. Where's
. Hang
, which
. Sit down

and just punctuation:

?
’

Each string is assigned to one of 100 functional linguistic-rhetorical categories, called ‘Language Action Types’ or LATs. Each LAT attempts to capture a set of words or phrases which the designers of Docuscope felt work together to produce the same functional effect in the reader of a text. For example, the string ‘I’ is assigned to the LAT ‘First Person’, which tags words that have a straightforward first person reference. Other words tagged in this LAT include: ‘my’, ‘myself’, ‘I’ll’, ‘mine’, ‘me’. However, Docuscope is functionally, rather than formally driven. It attempts to tag effects on the reader, not objective formal categories, so not all first person forms are tagged with this LAT. A set of words and phrases such as ‘methinks’, ‘I have’, ‘I would have’, ‘my daughter’, ‘my thoughts’, ‘from me’ are tagged in a separate, but related LAT, called ‘First Person Interior’. This LAT is designed to pick up words and phrases that reveal the interiority of the speaking self to the reader: a heightened degree of first person.

LATs associated with first person effects are relatively intuitive. Other LATs show the complexity, and subtlety, of the textual model Docuscope instantiates. For example, ‘here’ is tagged as part of the LAT ‘Spatial Relations’, which seeks to identify textual elements that specify space: ‘here’, ‘aground’, ‘aboard’, ‘over’, ‘tides’, furlongs’, ‘acre’. ‘Master’ is part of the ‘Person Property’ LAT, which collects terms for social and familial roles (‘Boatswain’, ‘mariners’, ‘counsellor’, ‘drunkards’, ‘brother’, ‘father’). Although it is a simple string-matching program, with no grammatical parsing capabilities, Docuscope uses context and punctuation to make likely distinctions between uses of similar strings.

So 'master' on its own is 'Person Property', but 'the master' is classed as 'Commonplace Authorities', a LAT which attempts to identify appeals to public sources of authority ('command', 'authority', 'warrant', 'prayers', 'god of', 'master of'). Similarly, 'what' immediately after a punctuation mark is tagged with the LAT 'Question' (as are individual question marks), and '. Hang' and '. Sit down' use the clue of punctuation to assign the verbs to 'Imperative'. 'Which' coming after a comma is tagged as 'Aside' (which includes unrestricted relative clauses) rather than 'Question'. We discuss other LATs in more detail below.

As we have said, Docuscope works by counting strings, which it assigns to LATs. It outputs the frequencies of each LAT in a corpus of texts as a csv (comma separated variable) file, which can be read by spreadsheet and statistical programs. Following standard statistical procedure, the frequencies of LATs are 'normalised': that is, every value is expressed as per a standard number of words to allow comparison between texts of different lengths. There are 100 LATs, and we have been investigating the 36 first folio plays, so the csv file Doscuscope generates for Shakespeare consists of 36 rows of data with 100 columns. The file is illustrated in Figure 1.

The complete csv file has 3,600 data points (100 values for each of 36 plays). How do we begin to make sense of this? We could start reading the csv file, looking for high and low points, but this would clearly be very time-consuming, and a human reader would likely miss many interesting comparisons and associations (do plays high in LAT X always have low values for LAT Y? Are Tragedies characterised by greater use of LAT W than Comedies?). To get at the relationships and patterns within this data, we use various visualisation techniques. An important point to make about visualisation, and indeed most forms of statistical analysis, is that it involves a *reduction* in the complexity of the information offered. The csv file, both in its form, and the amount of content it has, is unhelpful to human modes of understanding. It is hard for us to read, because we are not good at reading numbers, and it is hard for us to take in because there is too much information.

Humans may be bad at reading 36 x 100 tables of numbers, but we are good at taking in visual patterns, even quite complex ones, so it makes sense to represent the information in the csv file visually in some way. There are many ways of doing this: bar and line graphs; pie charts, and others. The aim, whatever visualisation we use, is to identify

patterns in the data we would not see if presented to us in csv form. The patterns are there in the csv file: and they are easily identified by computers using mathematics (computers being very good at reading csv files, and generally pretty bad at visual pattern recognition). There are many possible statistical patterns in such rich data – and not all of them will be interesting to us as literary scholars.

Figure 1: Partial view of the Shakespeare csv file generated by Docuscope

So far in our investigation of this data, we have been interested in comparison between plays and groups of plays. Our questions have been something like, 'Given what Docuscope counts, how similar, or different, is this play to other plays?', 'Do the differences and similarities Docuscope detects make sense in terms of the divisions literary scholars make within Shakespeare's work?' Our initial published work has concentrated on genre: do the things Docuscope counts pattern in accordance with genre in interesting ways? It turns out that they do. In this paper we shift our focus from genre to chronology, and we broaden our question from 'Does Docuscope recognise these categories identified by criticism (Comedy, History, Tragedy)?' to 'Does Docuscope recognise patterns of chronology in the corpus; are there groupings of plays that can be compared to the groupings made in literary history?'

Figure 2 is a visualisation which attempts to give us a first take on this question. It was generated from the csv file excerpted in Figure 1 by LATtice, a program developed by Dr Anupam Basu.[8]

[8] LATtice is described in, and can be downloaded from, the following post by Anupam Basu at our blog, winedarksea.org: Anupam Basu, "Visualising Linguistic Variation with LATtice", November 29, 2011, http://winedarksea.org/?p=1285.

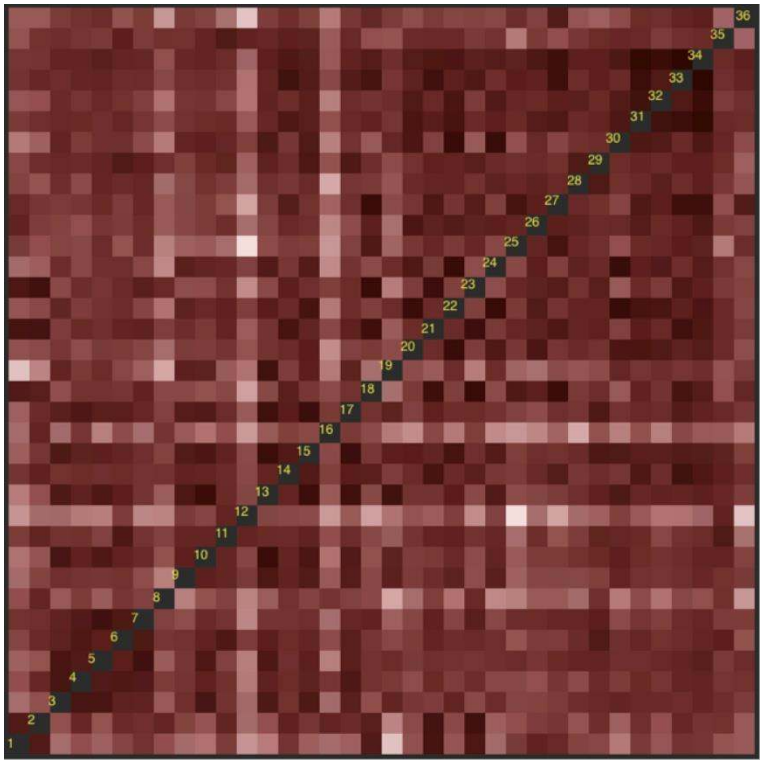

Figure 2: LATtice heatmap derived from the csv file excerpted in Figure 1

Having transformed 100 dimensions into two, we then transform distance into colour, representing difference by hue. The darker the square, the more similar the plays are (the closer they are in space); difference (distance) is indicated by increasing lightness. The square arranges the plays from the bottom-left corner in ascending chronological order, and this order is repeated from right to left along the horizontal base. So the 'first' square, at the extreme bottom left represents the earliest play being compared to itself. Hence the square is black – and hence the diagonal of black squares running across the main square as each play is compared to itself and found to be identical. The order of plays is given in Table 1.

Table 1: order of plays (from bottom left square upwards)

1 *The Two Gentlemen of Verona* (1590-1)
2 *The Taming of the Shrew* (1590-1)
3 *2 Henry VI* (1591)
4 *3 Henry VI* (1591)
5 *1 Henry VI* (1592)
6 *Titus Andronicus* (1592)
7 *Richard III* (1592-3)
8 *The Comedy of Errors* (1594)
9 *Love's Labour's Lost* (1594-5)
10 *Richard II* (1595)
11 *Romeo and Juliet* (1595)
12 *A Midsummer Night's Dream* (1595)
13 *King John* (1596)
14 *The Merchant of Venice* (1596-7)
15 *1 Henry IV* (1596-7)
16 *The Merry Wives of Windsor* (1597-8)
17 *2 Henry IV* (1597-8)
18 *Much Ado About Nothing* (1598)
19 *Henry V* (1598-9)
20 *Julius Caesar* (1599)
21 *As You Like It* (1599-1600)
22 *Hamlet* (1600-1)
23 *Twelfth Night* (1601)
24 *Troilus and Cressida* (1602)
25 *Measure for Measure* (1603)
26 *Othello* (1603-4)
27 *All's Well That Ends Well* (1604-5)
28 *Timon of Athens* (1605)
29 *King Lear* (1605-6)
30 *Macbeth* (1606)
31 *Antony and Cleopatra* (1606)
32 *Coriolanus* (1608)
33 *The Winter's Tale* (1609)
34 *Cymbeline* (1610)
35 *The Tempest* (1611)
36 *Henry VIII* (1613)

There are several caveats to be noted about this procedure. Heatmaps are intended to give a quick overview of a data set. They are good at this, but they give this overview at the cost of complexity: 100 differences summarised as one for each square. Heatmaps are quick to read; but human perception is not precise, and colour perception is highly relative. Depending on their surrounding squares, individual squares can appear lighter or darker than they really are. In addition, we need to remember that this heatmap uses just one of several chronologies for Shakespeare's works.[9] We need to use one, of course, but it may not be completely correct. Finally, we need to remember that as humans, especially as literary scholars, we are trained to find patterns and stories in anything we see. If we presented you with a heatmap of the plays in a random order, you would see patterns in that, and would construct narratives about the chronology of Shakespeare's career.

So what can we see in the heatmap? Our suggestion is that the heatmap offers us a rather different view of periodization in Shakespeare than has been traditional.

It seems to us that there are three groups of similar plays across Shakespeare's career (which runs chronologically from bottom left to top right). These groups are marked off in Figure 3. The first group, at lower left, consists mainly of the early histories: *1 Henry VI, 2 Henry VI, 3 Henry VI, Richard III*, and *Titus Andronicus*. To the naked eye, this group looks the darkest, or most self-consistent of all three. The next group, slightly above and to the right, consists of a more generically mixed group of plays, including later histories, a tragedy, and several comedies: *Love's Labour's Lost, Richard II, Romeo and Juliet, King John, The Merchant of Venice, 1 Henry IV* (it is tempting to alter the Wells-Taylor ordering slightly here to include *2 Henry IV*, which lies just beyond the group). Bisecting the square formed by these plays are the lines that represent *A Midsummer Night's Dream* – lines of strikingly light coloured squares. The final group is the largest, beginning with *Julius Caesar*, and containing *As You Like It, Hamlet, Twelfth Night,*

9 For this paper we have used the Wells-Taylor chronology, established as part of the Oxford Shakespeare project, and fully laid out in Stanley Wells and Gary Taylor, with John Jowett and William Montgomery, *William Shakespeare: A Textual Companion*, Oxford, Clarendon Press, 1987, p. 69-134. Where Wells-Taylor assign a period to a play (e.g. 1592-3), we take the early date; for the ordering of plays given the same, or overlapping dates, we follow the ordering used in the *Textual Companion*. Note that the csv file downloaded with LATtice (footnote 8) uses a different ordering to Wells-Taylor: users can easily alter this by reordering the plays in the csv file.

Troilus and Cressida, Measure for Measure, Othello, All's Well That Ends Well, Timon of Athens, King Lear, Macbeth, Antony and Cleopatra, Coriolanus, The Winter's Tale, and *Cymbeline*. Although they are included in most accounts of the Late plays, *The Tempest* and *Henry VIII* are notably lighter than the main block of later plays, indicating that they differ linguistically from them (at least in terms of what is being counted here).

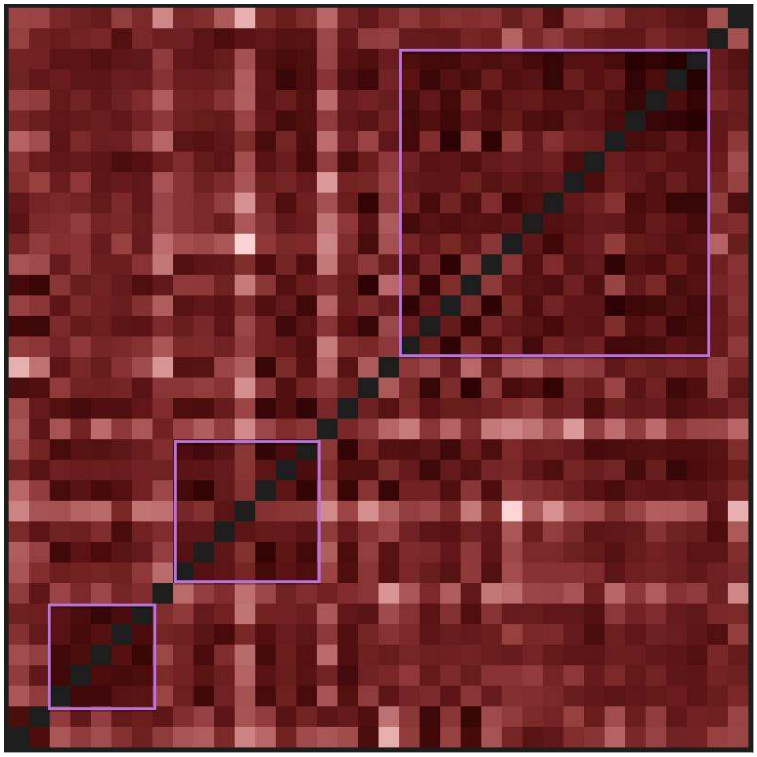

Figure 3: Groups of similar plays across Shakespeare's career

When we first looked at this heatmap, we found the large upper right square striking and surprising. We have worked on the late plays before, but in that work we restricted ourselves to the small group of plays identified by metrical analysis as Late. Critics have frequently identified these plays as highly characteristic in language and theme,

though they have also noted the presence of 'Late' linguistic features and themes in earlier plays.[10] The heatmap presents us with a much larger group of plays we will term 'later', as in 'the later plays' – and visually at least implies that Shakespeare gets more self-consistent (darker) as he matures.

But this 'result' is not an end point. There may be a statistical grouping here, but is this an *interesting* grouping? Does it open up new questions about Shakespeare's developing style, or offer new interpretations of the plays? To answer this, we need to look in more detail at the precise features that underlie the broad shadings of the heatmap, drilling down into the csv file to extract the LATs which make the later plays look self-consistent. Can we *explain* the shifts in frequency that are certainly there in ways that make sense to literary scholarship, or are we simply picking up a consistent drift in a set of linguistic features with no coherent stylistic outcome? When we shift our focus to individual LATs, we find that this increased similarity (the darkened square) is due to a range of LATs shifting in frequency together, one set rising over Shakespeare's career, another set decreasing. For reasons of space, we will consider just three LATs here: there is much more analysis to be done in terms of explicating the language of the later plays, and this paper should be considered an initial experiment.

We will begin with two Docuscope LATs which show a trend of decreasing frequency over Shakespeare's career. Figures 4 and 5 plot the frequencies of these LATs in each play arranged in chronological order. The normalised frequencies are shown on the vertical axis, while the play names are shown along the horizontal axis. We have added a linear trend line to each chart, using the in-built function in Excel (this can be found by following Chart > Chart Layout > Trendline).

Figure 4 shows the frequency of the LAT 'Sense Object' in each of the folio plays. We will show what this LAT does linguistically in a

[10] For an overview of accounts of the late style, see: Brian Vickers, "Approaching Shakespeare's late style", *Early Modern Literary Studies*, 13.3, January, 2008, 6.1-26. For the argument that elements of the later style appear earlier in Shakespeare's career, see Russ McDonald, *Shakespeare's Late Style*, Cambridge, CUP, 2006, p. 42-76. Gordon McMullan has challenged many of the assumptions that surround the notion of 'lateness' in artistic production: Gordon McMullan, *Shakespeare and the Idea of Late Writing: Authorship in the Proximity of Death*, Cambridge, CUP , 2007; Gordon McMullan, "What is a 'late play'?", in Catherine Alexander, ed., *The Cambridge Companion to Shakespeare's Last Plays*, Cambridge, CUP, 2009, p. 5-28.

moment, but first, we will discuss the graph, and the trending pattern. If you look at Figure 4, you should be able to see that the trend line is down over Shakespeare's career. It begins at around 3.7, and declines to end at almost 3.0. However, when we look at the individual results, we see that there certainly is not a one-to-one relationship between chronology and frequency of 'Sense Object': the play with the highest frequency of this LAT, *A Midsummer Night's Dream*, is relatively early in Shakespeare's career, and the first play (*The Two Gentlemen of Verona*) is very low. However, what we are interested in here is the overall trend, which is clearly down – and if we examine the individual results more closely, we see that of the nine plays which have a frequency score of less than 3.0 for 'Sense Object', seven come in the second half of Shakespeare's career, while another three in this group only just get over the 3.0 score. The trend line has thus identified an average decline in the use of this feature in the later part of Shakespeare's career, possibly with a beginning in *Much Ado About Nothing*. (And we might also note that the highest scoring 'later' play is *The Tempest*, which the heatmap has already signalled as not being typical of the later plays.)

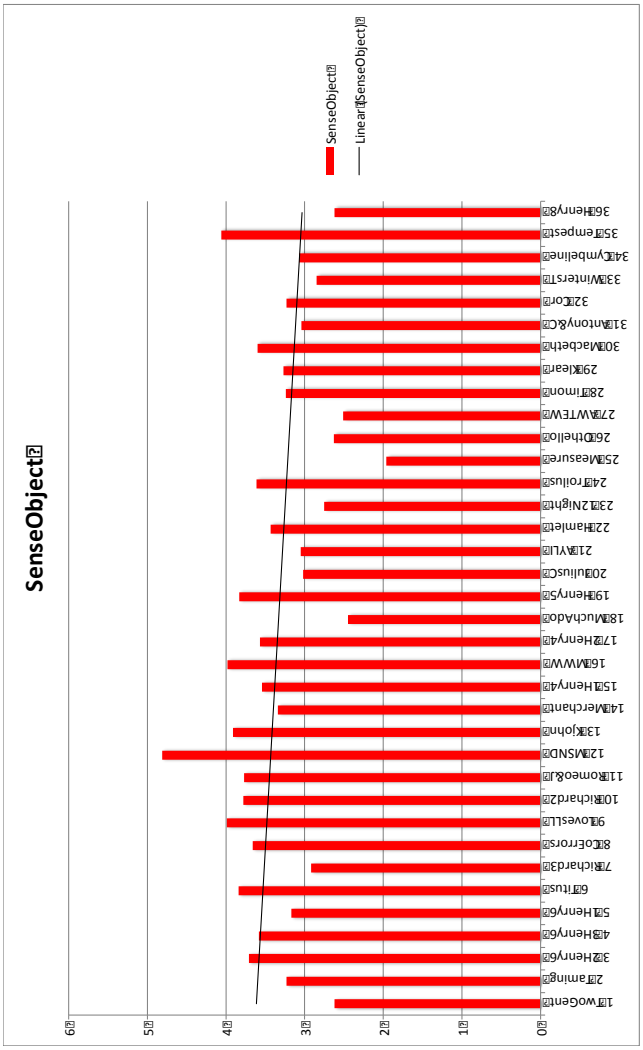

Figure 4: 'Sense Object' frequencies over Shakespeare's career

Figure 5 shows a similar declining trend, this time in the LAT 'Sense Property'. Here the decline is from c. 1.4 to just above 1.0, and again there are exceptions (*The Tempest* is higher than its neighbours; *The Two Gentlemen of Verona, 3 Henry VI,* and *The Comedy of Errors* are low). But once again, we can see that no play before *Much Ado About Nothing* dips below 1.0, while eight plays from the later period do so, including *Much Ado About Nothing* (and here we can see that adding in the extra detail of individual LAT frequencies, rather than using the massively reduced and averaged information of the heat-map, is suggesting that the later group might be extended even earlier in Shakespeare's career).

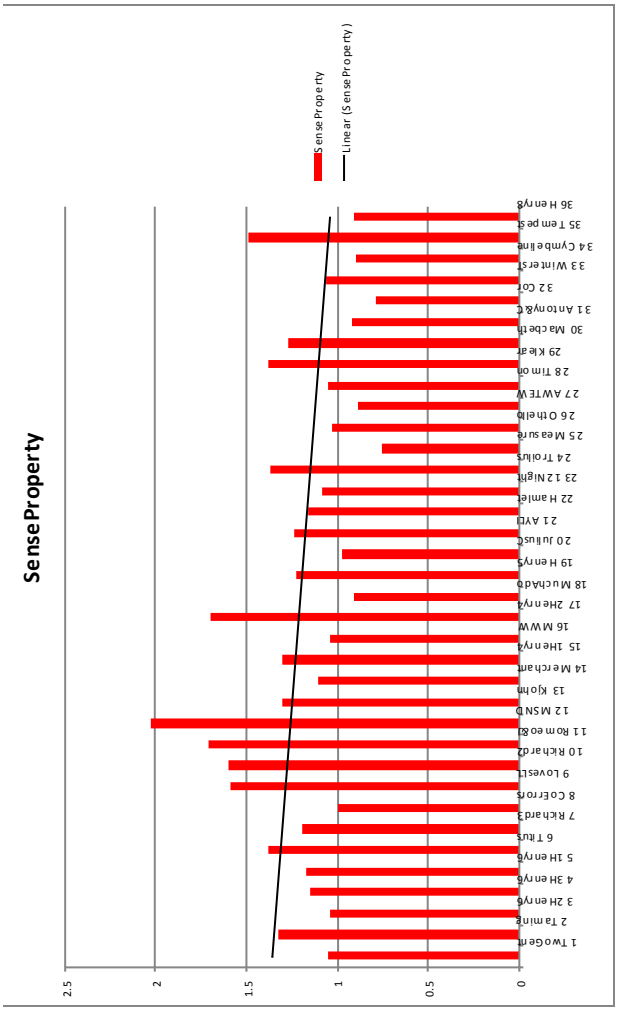

Figure 5: 'Sense Property' frequencies over Shakespeare's career

So those are the statistics. The question now is, 'Is this an interesting finding?' – can we as literary critics make any sense of it? In order to do this, we need to look at what these LATs actually do in texts. 'Sense Object' and 'Sense Property', as their names suggest, are closely associated LATs. 'Sense Object' mainly tags concrete nouns, while 'Sense Property' tags the adjectives and participles that describe the properties of those nouns. Thus if we look at results from *The Tempest*, unusually rich in both LATs for a later play, we find the following words tagged as 'Sense Object' in Prospero's narration of being cast away at sea (shaded yellow):

PROSPERO. [...] In few, they hurried us aboard a bark,
Bore us some leagues to sea, where they prepared
A rotten carcass of a boat, not rigged,
Nor tackle, sail, nor mast - the very rats
Instinctively have quit it. There they hoist us,
To cry to th' sea that roared to us, to sigh
To th'winds, whose pity, sighing back again,
Did us but loving wrong.

MIRANDA.
Alack, what trouble
Was I then to you!

PROSPERO.
O, a cherubin
Thou wast that did preserve me. Thou didst smile,
Infusèd with a fortitude from heaven,
When I have decked the sea with drops full salt,
Under my burden groaned; which raised in me
An undergoing stomach, to bear up
Against what should ensue.

MIRANDA. How came we ashore?

PROSPERO. By Providence divine.
Some food we had, and some fresh water, that
A noble Neapolitan, Gonzalo,
Out of his charity – who being then appointed
Master of this design - did give us; with
Rich garments, linens, stuffs, and necessaries
Which since have steaded much. So, of his gentleness,
Knowing I loved my books, he furnished me
From mine own library with volumes that
I prize above my dukedom.

The Tempest, I.ii.144-169

You will note that not all nouns are tagged: for example, 'tackle' and 'sail' in line 147. This is because Docuscope is a simple string-matching program. It has no part of speech analysis, so strings are allocated to what the designers saw as their primary function. In this case, 'tackle' and 'sail' are assumed to be verbs ('she tackled the most difficult question', 'she will sail single-handed') and are assigned to the LAT 'Motion'.

Here are the 'Sense Property' tokens from Ariel's speech describing his raising of the storm (again, shaded yellow):

ARIEL. [...] I boarded the King's ship. Now on the beak,
Now in the waste, the deck, in every cabin,
I flamed amazement. Sometime I'd divide,
And burn in many places; on the top-mast,
The yards, and bowsprit, would I flame distinctly;
Then meet and join. Jove's lightning, the precursors
O'th' dreadful thunderclaps, more momentary
And sight-outrunning were not. The fire and cracks
Of sulphurous roaring the most mighty Neptune
Seem to besiege, and make his bold waves tremble,
Yea, his dread trident shake.

PROSPERO.
My brave spirit!
Who was so firm, so constant, that this coil
Would not infect his reason?

ARIEL.
Not a soul
But felt a fever of the mad, and played
Some tricks of desperation. All but mariners
Plunged in the foaming brine and quit the vessel,
Then all afire with me. The King's son Ferdinand,
With hair upstaring - then like reeds, not hair -
Was the first man that leaped; cried, 'Hell is empty,
And all the devils are here.'

PROSPERO.
Why, that's my spirit!
But was not this nigh shore?

ARIEL.
Close by, my master.

PROSPERO. But are they, Ariel, safe?

ARIEL.
Not a hair perished.

> On their sustaining garments not a blemish,
> But fresher than before. And, as thou bad'st me,
> In troops I have dispersed them 'bout the isle.
> The King's son have I landed by himself,
> Whom I left cooling of the air with sighs
> In an odd angle of the isle, and sitting,
> His arms in this sad knot.
>
> *The Tempest*, I.ii.197-225

Again, not all adjectives are tagged as 'Sense Property' – for example 'dreadful' (203), 'fresher' (220), and 'sad' (225). In this case, this is not because Docuscope does not categorise them as adjectives: it does. Rather this is because Docuscope has many categories, and strings can only be a member of one. In this case, 'dreadful' has been assigned to the LAT 'Fear', 'fresher' to the LAT 'Positive Emotion', and 'sad' to the LAT 'Sadness'. The overlapping claims of formal and semantic analysis are demonstrated here: Docuscope's categories are informed by formal analysis ('Is this string an adjective?'), but ultimately are determined by function ('Whatever this string is formally, what is it doing functionally in the text and to the reader?'). Certainly, a linguist could object at the mixing of formal categories in Docuscope LATs, and also at the requirement that strings can belong to only one LAT (since we know that linguistic features are polysemic). And any user of Docuscope could argue with the (ultimately often subjective) assignment of strings to semantically driven LATs: should 'fresher' be in 'Positive Emotion' or 'Sense Property'? But the great strength of Docuscope is its scope (it tags upwards of 70% of strings in most texts across 100 features) and its consistency: it treats every text the same, and its miss-taggings are consistent.

It is clear that 'Sense Object' and 'Sense Property' function to communicate a vivid sense of the material world external to the speaker and audience (even if that world is sometimes ethereal or lacking solidity – Cherubin, flamed, sulphurous, afire). In our previous work, we have noted the importance of 'Sense Object' and 'Sense Property' in the Histories, which emerge as deeply concerned with the denotation of the external world. The downward trend in the use of these two LATs over Shakespeare's career suggests that his writing becomes less engaged with the external world, less insistent on naming and describing. It would be wrong to overstate this: this is a trend, not a sudden abandonment of nouns and adjectives; but a suggestive group of plays

at the forefront of this trend emerges from Figure 4: *Much Ado, Measure for Measure, Othello, All's Well That Ends Well*. We would suggest that further exploration of this as a possible marker of a 'later' style in Shakespeare would be productive.

At the same time as he lowers his use of these LATs associated with the denotation of the external world, Shakespeare increases his use of 'Person Pronoun', a slightly unusual Docuscope category in that it is almost entirely formal. This LAT tags third person reference, typically the third person pronouns 'she', 'her', 'he', 'him', and 'their', as well as some relative pronouns such as 'whose'. Figure 6 shows the results for this LAT across the chronology of Shakespeare's plays: the average goes up from around 1.6 to just under 2.0, and of the eight plays which have a frequency score of over 2.0, seven of them are in the later period. Of the five plays with a score of under 1.5, four are in the early period – and the one later play with a score this low is *The Tempest*, consistent in its linguistic departures from most of the later trends.

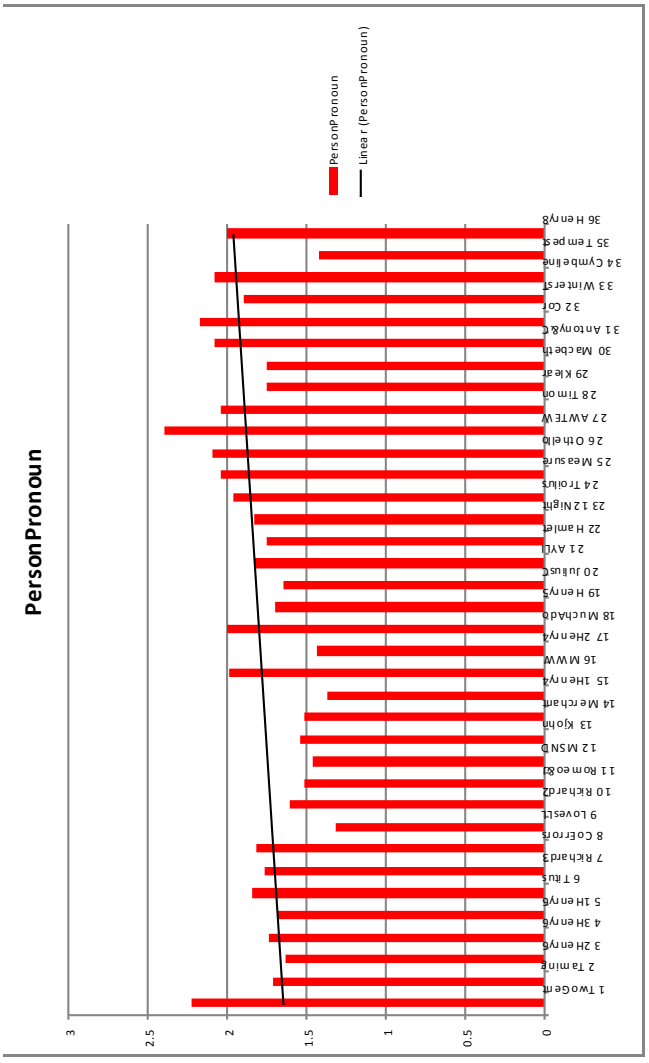

Figure 6: 'Person pronoun' frequencies over Shakespeare's career

Here is the opening of *All's Well That Ends Well*, the highest scoring play for this LAT, with 'Person Pronoun' tokens shaded yellow:

COUNTESS. In delivering my son from me I bury a second husband.

BERTRAM. And I in going, madam, weep o'er my father's death anew; but I must attend his majesty's command, to whom I am now in ward, evermore in subjection.

LAFEU. You shall find of the King a husband, madam; you, sir, a father. He that so generally is at all times good must of necessity hold his virtue to you, whose worthiness would stir it up where it wanted rather than lack it where there is such abundance.

COUNTESS. What hope is there of his majesty's amendment?

LAFEU. He hath abandoned his physicians, madam, under whose practises he hath persecuted time with hope, and finds no other advantage in the process but only the losing of hope by time.

COUNTESS. This young gentlewoman had a father - O that 'had': how sad a passage 'tis! - whose skill was almost as great as his honesty; had it stretched so far, would have made nature immortal, and death should have play for lack of work. Would, for the King's sake, he were living. I think it would be the death of the King's disease.

LAFEU. How called you the man you speak of, madam?

COUNTESS. He was famous, sir, in his profession, and it was his great right to be so: Gérard de Narbonne.

LAFEU. He was excellent indeed, madam. The King very lately spoke of him, admiringly and mourningly. He was skilful enough to have lived still, if knowledge could be set up against mortality.

BERTRAM. What is it, my good lord, the King languishes of?

LAFEU. A fistula, my lord.

BERTRAM. I heard not of it before.

LAFEU. I would it were not notorious. - Was this gentlewoman the daughter of Gérard de Narbonne?

COUNTESS. His sole child, my lord, and bequeathed to my overlooking. I have those hopes of her good that her

 education promises; her dispositions she inherits, which makes fair gifts fairer - for where an unclean mind carries virtuous qualities, there commendations go with pity: they are virtues and traitors too. In her they are the better for their simpleness. She derives her honesty and achieves her goodness.

LAFEU. Your commendations, madam, get from her tears.

COUNTESS. 'Tis the best brine a maiden can season her praise in. The remembrance of her father never approaches her heart but the tyranny of her sorrows takes all livelihood from her cheek. - No more of this, Helen. Go to, no more, lest it be rather thought you affect a sorrow than to have–

All's Well That Ends Well, I.i.1-50

Note that not all third person pronouns are tagged: in line 6, 'he' is not tagged because it is included in a longer string, 'he that', which is counted as the LAT 'Commonplace Authority' (typically an appeal to external higher authority, such as the law, government, or God). Docuscope's algorithms work so that it only counts the longest possible string for each word. The same applies to 'his' in line 23, which is part of a longer string 'it was his', tagged as the LAT 'Biographical Time', which marks narratives of time.

What can we say about the possible effects of this increase in 'Person Pronoun' over Shakespeare's career? In particular, can we link this to the apparent recession of the physical world suggested by the decreasing use of 'Sense Object' and 'Sense Property'? At first sight, it would appear not. Pronouns, like nouns, tend to have real-world referents: it could be argued that they too are concerned with denoting the external world. So are we simply charting random drift here, with no literary interest? We suggest not. The effect of pronoun reference is different to that of full nouns, as used in 'Sense Object'. Pronouns depend for their reference, not on real-world objects and semantic knowledge, but on discourse-internal knowledge. If we write the word 'sea', it has reference for a reader because they know the linguistic conventions that give semantic weight to the word: they bring their external understanding of it to the text. But if we write, or speak the word 'she', then the precise reference of that word is determined by the context in which it is used: it is internal to the discourse in which it appears, and is entirely dependent on that context.

Pronoun reference is therefore different to denotational reference by nouns, because it relies on the immediate context, and the implicitly shared knowledge of speaker and hearer, rather than the community knowledge held in *langue*. Crucially for our argument here, a shift from nouns to pronouns means that the focus switches from the external world (denoted objectively by nouns) to a version of the external world viewed through the subjectivity of the speaker: 'sea' means what everyone agrees it to mean whenever and whoever says it; 'she' means whatever I intend it to mean at the point of utterance.

A further aspect of pronoun reference is relevant to considering Shakespeare's later style. In writing, pronouns normally appear only after the full nouns to which they refer have been used. We mention 'Elizabeth' in the first sentence, and then switch to 'she' subsequently, knowing that readers have the context to supply the referent. This process is known as 'pronoun replacement'. Speech, however, tends to function slightly differently, in that shared contextual knowledge is implicit between speaker and hearer: 'she' is in the room, or has just left, so naming her is unnecessary. This shared context means that speech normally has a higher frequency of pronouns than writing, while writing has a higher frequency of full noun phrases to avoid ambiguity.

The opening scene of *All's Well* has a high frequency of pronoun usage, with full names delayed, contravening the strict expectations of pronoun replacement (and arguably the needs of the audience): 'Gérard de Narbonne' is named at line 23, but has been repeatedly referred to previously ('a father', 'his', 'he', 'the man you speak of', 'he'); Helen is named, almost as an afterthought, only following repeated pronoun reference ('this gentlewoman', 'the daughter of Gérard de Narbonne', 'His sole child', 'her', 'her', 'her', 'she', 'her', 'she', 'her', 'her', 'her', 'her', 'her', 'her', 'her', 'her'; and then finally: 'No more of this, Helen'). Of course, while this contravenes what we would expect in formal writing, it is exactly what we expect in speech, and it is one of the stylistic features that makes the opening of the play seem casual: the conversation flows quickly and naturally, and we have the impression of being thrust *in medias res*.

Ultimately, this scene is 'well-behaved', in that all the pronouns are given referents, and delaying the names both pulls us into the action, and makes the language seem realistic. It is, however, not hard to find passages in later Shakespeare where the density of pronoun

replacement, and indeed the use of multiple instances of the same pronoun form to refer to different referents, cause ambiguity. *Macbeth* is characterised by such murky language - note the shifting and uncertain referents of 'what', 'that' and 'it' in Lady Macbeth's character assessment of her husband:

> Thou wouldst be great;
> Art not without ambition, but without
> The illness should attend it: what thou wouldst highly,
> That wouldst thou holily; wouldst not play false,
> And yet wouldst wrongly win; thou'dst have, great Glamis,
> That which cries, 'Thus thou must do,' if thou have it;
> And that which rather thou dost fear to do,
> Than wishest should be undone. (I.v.17-24)

Shakespeare's shift to a higher frequency of 'Person Pronoun' in his later work not only shifts the focus from the 'real' world to the perception of that world in the mind of the speaker, but also makes his later style look more like speech. Inevitably, this entails a danger of increased ambiguity, since it can become difficult to determine the referent of pronouns if it is not given, or is withheld. Tellingly, literary critics have often identified increased ambiguity and difficulty of reference in their responses to Shakespeare's later style.

Although our findings are preliminary, and limited by space here, we can already say that we have identified other LATs, the trends in which seem to share this shift away from real world denotation, and into a focus on the subjectivity of the speaker. Certainly, critics working non-quantitatively have suggested that something similar to this is going on, and we take encouragement from this. Our aim is not to overthrow traditional methods, but to work alongside them, and add to their evidence base.

Jonathan Hope
Strathclyde University, Glasgow

Michael Witmore
Folger Shakespeare Library, Washington D.C.