'*Après le déluge*, More Criticism': Philology, Literary History and Ancestral Reading in the Coming Post-Transcription World

Michael Witmore (Folger Shakespeare Library) and Jonathan Hope (Strathclyde University, Glasgow)

1 Early Modern Studies and the Coming Deluge of Data

'Work as if you lived in the early days of a better nation' is a motto found throughout the work of the Glaswegian artist and writer, Alasdair Gray – and it is one which might well be applied to those engaged in what has been called 'digital humanities'. The future offers the promise, by 2017 if the estimates of the Text Creation Partnership are correct, of not just the whole of printed renaissance drama, but every surviving printed book in English from the early Modern period, transcribed and available for searching and 'data mining' on the laptops of every academic, postgraduate, and undergraduate in the field. Whatever this nation will be, if nation is really the right word for the inhabitants of this post-deluge, information-rich world, it will have many members.

Even now, large sections of the drama are available – far more than the most assiduous readers can claim familiarity with (and this promise is also a threat to which we will return).[1] But material availability is simply the most basic step in research: the transcription rates of the TCP are impressive, as is the scale and ambition of the project, but they are merely providing data. It will be up to academics and their students to use the data, and this will require not just work, but relearning and re-theorising. It will also involve the field in a shift in research methods and teaching requirements.

As the full range of Renaissance printed material becomes available, our subject changes: we can ask different questions and we can answer the old ones with new, alien methods. Someone planning a book on the history of English genres in 1960 could legitimately have laid out a research plan that would have been equally legitimate in 1860, or 1760: a wide range of reading, certainly, but selective reading, and even more selective writing when it came to evidence. The narrative would arise out of the reading, but that reading could be only partial, so the claims made would have to be judged on rhetorical grounds: are they persuasive, seductive? No? How else could it be said?

Plan a book about the rise and fall of genres now, and it is possible – and arguably necessary – to 'read' everything, or at least everything that has been digitised. The narrative, if there is one, lies in the data and the claims stand or fall on the size of the sample, the statistical significance of the results, the care with which the procedures have been applied.

Franco Moretti has written such a book for the genres of the novel – and we are planning a study on the genres of English prose in the seventeenth century.[2] In the past, such works would have been 'magisterial' in the sense of broad surveys which constructed a grand narrative, judged on the quality of their prose and organization of materials. Now we are faced with something more like an epidemiology of literary populations, light sluicing across a map where the virus has passed. We are required not simply to identify a number of exemplary texts, but also to take account of all the evidence–every case, every death, every recovery.

2 Pedagogy and Research

What is brave about this new world is the ridiculous availability of data; to live in it, our students will need different skills, and we will have to change the way we prepare them. Interestingly, we are finding that the gap between pedagogy and research is narrowing: as we learn new methods, our students fall behind; and the nature of the research procedures mean undergraduates and post–graduates can make genuine contributions.

A literature student in 2020 will, as a matter of course, want to run basic concordance and other procedures on their assigned texts – their tutors/advisors (in advanced instruction at least) will expect them to run log likelihood tests before writing, for example, on 'love' in Shakespeare (and how long will it be before 'log likelihood' can be used in a context like this without explanation, as holds now for 'Foulcauldian', 'tropological', or 'Henslowe'? We provide an explanation later in this paper.). In order to run these tests, they will need to know about concordance programs and basic statistics, but they will also need to know about text curation. Metadata is decisive: author, title, date, genre – without this, electronic texts are literally useless. One might as well try to study transcripts of ocean temperatures around the globe without knowing where the readings were taken. Students will also need to know about text preparation: does their source text have line numbers, act/scene numbers, speech assignments which might be counted by software as if they were 'real' parts of the text? Are they not 'real' parts of the text? Our students will also need to know something of the theory of textuality, and a lot about renaissance printing and book manufacture – to say nothing of the processes of modernisation.

To say something of textual modernisation, our students will need to understand why it is a good thing (it makes searching texts easier) and why it is a bad thing (it is not a mechanical process divorced from symptomatic reading). Here, they will encounter a substantial theoretical problem. As Margreta de Grazia has argued, given the absence of dictionaries and stable spelling systems in the Early Modern period, 'words' as we understand them may not really exist.[3] This is an uncomfortable proposition for a process that seeks to 'resolve' Early Modern variation and ambiguity into a stability imposed on the language after the period – neutralising a key linguistic process which derives meaning through context (which we could call the problematic of words) in favour of one which locates meaning outside the text, in the dictionary.[4]

We will also need to get used to living in a discipline where knowledge is cumulative (you need to understand one thing before moving on to the next) and advancing (things turn out to have been wrong, and are replaced by things that are less wrong). At present, literary studies are not cumulative in the sense that you can read pretty much in any order, using any one or more of a whole range of theoretical approaches. You don't *need* to read *Hamlet* before you read *Waiting for Godot*, and you don't *need* to have read Marxist theory before you can say something interesting about either (though it might be handy). *Après le déluge*, however, if you don't understand how concordances work, you won't understand the limitations of what they appear to show. And if you don't understand some basic statistics, you might find yourself making apparently common–sense, but actually false, claims about Shakespeare's vocabulary - that he had a larger vocabulary than other writers, for example, or that he invents words more frequently.

The question of Shakespeare's vocabulary offers an excellent example of the ways our discipline will change. It is one of the few areas of literary studies where 'traditional' literary critics have been happy to cite quantitative evidence in support of literary claims:

it is not hard to find writers on Shakespeare, and English, who refer to the 'facts' of Shakespeare's huge vocabulary and stunning ability to coin new words.

We will begin with the first of these 'facts'. Perhaps its most recent iteration comes in David Crystal's popular textbook on Shakespeare's language, in a chapter in which he attempts to debunk other 'myths' about Shakespeare's words.[5] On the face of it, the claim that Shakespeare had a larger vocabulary than his contemporaries looks convincing: Shakespeare uses about 20,500 different words, while Jonson lags behind on c. 19,000, with Middleton bringing up a suitably ignominious rear at c. 14,000. Score one for the genius of Shakespeare. Except not: common sense is often not very good at dealing with statistics, and in this case we need to allow for the *amount* of text by each writer that survives. Shakespeare's vocabulary looks bigger because more plays by him survive. He therefore gave himself more opportunity to use different words, rather as a batsman in cricket has the opportunity to score more runs if he plays more games. Instead of comparing the raw totals of words, we need to compare the rate at which these writers use words they haven't used before (just as in cricket, batsmen are compared on their average score per innings, rather than the raw total of career runs).

Once we look at this rate, we find that Shakespeare is no different from Jonson or Middleton in the size of his vocabulary. If Jonson or Middleton had written as many plays as Shakespeare, they would have ended-up using a similar total of different words. The year 2011 is likely to go down in literary history as the year several scholars, initially independently, but then collaboratively, established the size of Shakespeare's vocabulary and demonstrated that it is not larger than that of other writers.[6]

Related claims about Shakespeare's supposed linguistic fecundity have been challenged since the 1980s when the work of Jürgen Schäffer showed that the *OED* massively overestimates Shakespeare's 'invention' of words. Shakespeare features as the first citation for hundreds of words, and new uses of old words, not because he was actually the first to use them, but because his work was searched more thoroughly than other writers'.[7] Ante–datings to these first–citations are found very frequently: and the availability of the complete corpus of printed Renaissance texts will make them even easier to find. We can also say that claims about Shakespeare's supposedly exceptional fecundity have been undermined by the observation (by Schäffer and Nevalainen)[8] that *everyone* was making up words in the Renaissance – and (by Spevack)[9] that Shakespeare seems highly averse to the most salient source of new words: Latin. This discovery helps us put our finger, perhaps, on the different verbal texture of Shakespeare's writing from that of his peers, such as Jonson.

These are all findings based on quantitative studies that refine and correct earlier studies – these earlier studies either being based on poor assumptions about the data (that *OED* sampling treated all writers equally) or a failure to contextualise an apparently striking result in its wider population. Thus the procedural point that *OED* 'first–citations' are still provisional. The *OED* very specifically makes no claim about 'first–use' on the grounds that (i) not all extant printed texts were exhaustively searched (though they soon will be); (ii) that very few manuscripts were searched; and (iii) that no Early Modern conversations were recorded and transcribed. Even after we have the entire printed corpus available, we will not be able to claim as a 'first–use' the first appearance of a word in print: there may be earlier manuscript uses, or the word may first have been used in speech, not to mention the possibility that the word was first used in one of the thousands of books or pamphlets which do not survive.

As the work of Craig and Elliott and Valenza's suggests, digital and quantitative research in the humanities is carried out on a different paradigm from 'traditional' literary critical work. Instead of individuals working for long periods in isolation, publishing what they hope is 'the last word' on their topic, we have groups of scholars publishing simultaneously on similar topics, exchanging work in advance, and refining techniques through genuine exchange on blogs and websites, and via small conferences. In our own work, we have blogged about ongoing refinements to our techniques and on source texts: published work improves on previously published work because we improve our techniques and source texts, as well as our interpretation of the results. Literary scholars have long prefaced conference papers with the caveat that the work they present is 'in progress' – with the implication that the published for of the work will offer some more closed, authoritative version of the work. But in the sciences, such disclaimers go without saying: all work is 'in progress'; all findings provisional; all techniques open to refinement.

The post-deluge datasphere of criticism offers a very different kind of pedagogic experience for students and instructors: it is now possible for undergraduate students to make genuine discoveries about Shakespeare's language. In an undergraduate class taught by Hope on 'Shakespeare and Language' (QQ708, Strathclyde University 2010), students with little or no linguistic or computing training have been using Martin Mueller et al.'s excellent, and very user–friendly program Wordhoard to run log–likelihood analyses of the plays they have been individually assigned to study.[11] Log–likelihood is a more sophisticated analysis of word frequency: rather than simply tell you the most frequent content words in a play (a reliable method of confirming to yourself obvious things you already knew about the topic of a play), log–likelihood tests a play against the whole of the rest of Shakespeare, producing a list of those words Shakespeare uses more *and less* frequently than expected in the analysed play. This can produce surprising results which form jumping–off points for traditional literary enquiry.

It is hardly noteworthy, for example, that the word 'monster' is relatively far more frequent in *The Tempest* than in the rest of Shakespeare, but it surely is surprising and fascinating, that the word 'love' is used far *less* frequently in this play than in Shakespeare as a whole. As Fiona Paterson, the student who made this discovery, pointed out, 'Isn't *The Tempest* supposed to be a Romance?' Wordhoard is functioning in this situation as a discovery platform for students: it does not provide them with answers; it provides them with questions and problems they would not otherwise have come across. Without digital methods, another student, Mairi Fullerton Pegg, would not have discovered that the pronouns 'he' and 'it' are significantly *less* frequent in *Midsummer Night's Dream* than in the rest of Shakespeare. Now it is up to her to go back to the text and try to explain her finding.

The skills of text and data management, as well as basic familiarity with some statistical notions being developed here, will, of course, be appreciated by potential employers – but there are larger intellectual issues of numeracy. When did it become almost a point of honour for Humanities students not to be even basically numerate? Surely anyone we expect to be able to write about Shakespeare should be able to think about simple statistics and what they tell us? Humanists, averse to statistics, seem nevertheless easily dazzled by diagrams and visualizations. This is partly understandable: visualizations make obvious a complexity that we intuitively recognize as the subject and challenge of our work. But we and our students will need to be considerably more literate in the arts of visual comparison, being able, for example, to read a dendrogram or scatterplot when the situation demands it.

3 Practice and Theory

In addition to expanding our research and teaching, digital inquiry poses exciting theoretical questions for literary research. Like close reading, exploratory statistical analysis is partly an art, one that requires a certain scepticism and creativity. The inventor of this field of study, John Tukey, makes this point when he advises inquirers to delay hypothesis formulation and testing: sometimes it is more productive use descriptive statistics to invent new questions, in part by surveying patterns one did not initially think were important.[12]

Not all statistical tests are appropriate to all types of data, moreover, and it is all too easy to generate meaningless data sets with apparently astounding results by preparing the texts in the wrong way. We faced this problem early on in our research when we began to study the prevalence of certain types of concrete nouns and words describing the status of persons in our work with the text-tagging program Docuscope.[13] We learned, for example, that a statistical procedure known as Principal Component Analysis was capable of picking out a group of related plays – plays that exhibited certain common linguistic features but also failed to exhibit others in a common way – which we, as critics, recognized as Shakespeare's history plays.[14] For our analysis, we were using the Moby Shakespeare, itself based on the one-volume 1864 Globe Edition edited by William George Clark and William Aldis Wright. This edition, like many other modern editions, introduced standardized speech prefixes for each play. When we "counted" the plays, assigning various words and phrases to various rhetorical categories that are built into the program, we obtained results that seemed impossibly accurate with respect to the history plays. An incredible finding! Shakespeare did something remarkably distinctive when he wrote history plays, and this distinctive thing made the plays as a group visible through statistical analysis from a mile away.

Now, it turns out that Shakespeare really does do something consistent when he writes history plays, but the degree of distinction we were seeing was overstated. Because we had retained the speech prefixes in our source texts, they were being counted by the program, which was assigning them quite frequently to a category known as "Person Property," a collection of words and phrases that indicate the social status or profession of an individual. When these nineteenth century editors standardized the speech prefix of monarchs to "King," followed by the monarch's Christian name, Docuscope tallied up each occurrence. So of course, the history plays stood out like a sore thumb. They were covered with the unvoiced word "King" in the speech prefixes, and these were a tip-off or "tell" to the statistics that a certain type of play was being profiled. We decided to strip off the speech prefixes, stage directions and act/scene divisions and run the analysis again, having decided that for our purposes this material could be misleading. Arguably this was a sound decision, because we wanted and still want to find something in the linguistic texture of the spoken language of the plays that correlates with theatrical genre.

But it remains a motivated decision to identify the "linguistic substrate" of Shakespeare's plays with the spoken words in performance. What, exactly, is the difference between the apparatus of the printed play – speech prefixes, act and scene divisions, stage directions – and the language that is spoken by characters? This remains a good question. When we were forced to make this decision, however, we had to ask ourselves what we were really looking for. We wanted to identify the syntactic and rhetorical markers of genre – the ones that are lodged in the receptacle of individual sentences – rather than markers which attach to the play-text's characterization of speakers (such as the speech prefix

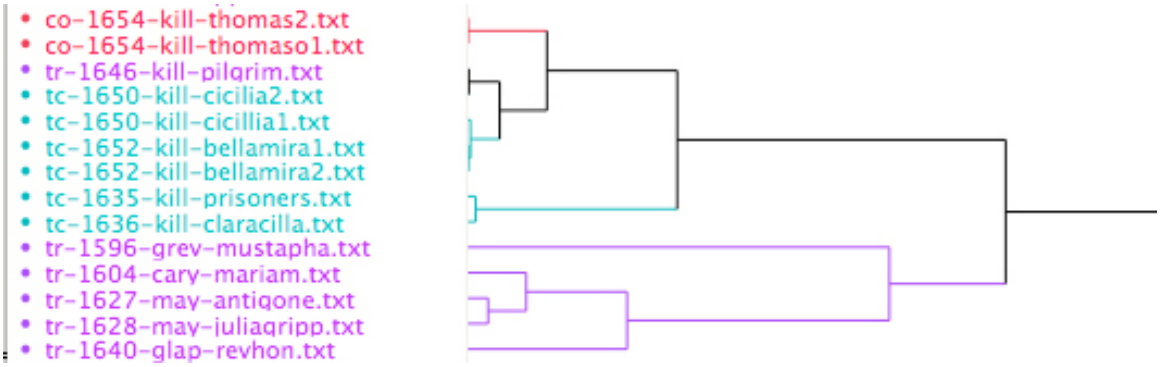
"King"). We were looking, that is, for something like a linguistic unconscious of genre, one that authors and readers cannot consciously attend to, but which nevertheless is the result of certain kinds of generic constraints on language use.

We had another illuminating moment when we attempted to work with digitized versions of early modern texts that had been subjected to different editorial procedures. Here we were attempting to identify differences between Shakespeare's dramatic writings and those of other early modern dramatists who were active between the years 1519-1659. Once again, we came up with a striking result that would have been easy to embrace because of its dramatic nature. Shakespeare, it appeared to us, seemed frequently to use a type of language that our program had identified under the category of "Language Reference," a type of language that refers reflexively to language itself. We decided to look more closely at the result. It turned out that the instance of Language Reference that was being tagged most often was the exclamation "O," which in the Globe Shakespeare had been systematically standardized from its variant form "Oh," which was used just as often in the rest of the collection of plays we were working with. The non-Shakespearean texts we were working with, on the other hand, had been modernized semi-algorithmically from the TCP original spelling source files. As a result, there was much more variation between O and Oh in the non-Shakespearean texts – a variation that had been artificially eliminated by the hand curation of the Globe Shakespeare text – which was making Shakespeare's texts, once again, highly visible.

But this result was misleading too. When we examined the result sceptically, returning as we always do to individual passages in the plays to see what exactly was being counted and which words or types of words were tipping the scales toward a particular result, we realized that what we were seeing was *editorial* rather than authorial difference in the grouping of our texts. (This continual return to the text is part of what makes digital analysis "iterative" in our conception of it.) Now, this result was encouraging in one respect: we had learned that any difference in editorial procedure used to create a group of texts, if consistent across only part of the whole group, will be visible statistically. Docuscope was "seeing" hand-editing rather than authorship. But the result also showed us that we needed to be absolutely consistent in our treatment of words that have variant spellings, which means the majority of early modern words. Because certain types of statistical analysis will detect *any* form of unevenly distributed variation, editorial preparation of texts (what we now call "pre-processing") is crucial feature of textual digital studies. Textual studies have always been sensitive to this phenomenon of orthographic variation and has, in effect, already paved the way for thinking about this issue. But digital inquiry into early modern texts has made the problem obvious in a new way, since all forms of iterative or algorithmic criticism rely on the power of consistent or standardized counting to obtain results.

Traditional literary scholars ought to welcome the implication of this: to do good digital work, you need to pay the closest possible attention to your texts, and their history. You need to know your subject area as well. During our work with this larger corpus of early modern plays (150 years worth), we showed a dendrogram of the entire group to one of our colleagues, Karen Britland. A dendrogram is designed to show degrees of similarity and dissimilarity in a population of items by ordering them into something like a tree diagram. Plays which appear closely linked to each other are similar linguistically - the deeper the separation, the less alike they are. The diagram was part of our exploratory work; it was ungainly, but it showed us that sometimes authorship is the decisive factor in grouping certain texts, while at other times genre or time of composition becomes

more important.[15] While many of the twigs of the dendrogram seemed intelligible to us, the following was puzzling:



Here you can see a list of plays on the left, tagged with their genre ('co' = comedy; 'tr' = tragedy; 'tc' = tragicomedy), date, author's name ('kill' = Killigrew) and abbreviated title. The tree-diagram on the right indicates similarity: note how all the plays by Thomas Killigrew are grouped together on one sub-branch. It is clear that the factor producing this upper branch is authorship. But we had no explanation for the lower branch grouping of apparently disparate plays. After we showed the diagram to Karen Britland, who is currently working on the literary and cultural milieu of Elizabath Cary, she pointed out what we should have already known: these texts are produced within two generations of the Sidney circle, and the similarities among them represent (in part) the preoccupations of that circle. Britland's reaction and ability to identify the pattern reinforced a lesson we had already learned in our work with Shakespeare: unless you know the literary-historical landscape of the texts you are studying, you will not be able to interpret the results of statistical studies of such texts. Score another point for traditional humanistic knowledge: the humanities produces crucial categories – what computer scientists called meta-data – which are the basis for interpreting results. No trained humanists, no metadata. And no results.[16]

There are also significant philosophical questions for digital inquiry, which require deliberate, abstract reflection. Perhaps within the next decade, it will be possible to have the whole corpus of surviving English printed books on your laptop. You will be able to call up every surviving instance of a particular word or phrase via a search that takes seconds. What, strictly speaking, will you be looking at when you look at the results of such a search? And what could, or should, you do with the results?

It could be said, for example, that you will be looking 'at' Early Modern Culture in some sense – a massively larger, more powerful version of the Arden–type footnotes that cite analogues and examples from other plays of the uses of selected words and phrases. Such a corpus can be the platform what Daniel Shore has recently called "plural reading," a way of scanning syntactical units within a corpus that is both close to and far away from the individual texts in question.[17] We can, as Shore has done, point to the first known occurrence of the subjunctive mood being used in counterfactual invocations of Jesus's ethical judgments: if Jesus were alive today, he would…. Such an occurrence suggests that a *certain way* of understanding the distance of the present from the past becomes possible at a given moment: we move from an occurrence to a state of mind or a perceptual possibility. But of course, we cannot know whether a cognate discourse – casuistry, for example – introduces another idiom for counterfactual thinking that produces the same deliberative effect with different syntactical units.[18] We cannot know,

that is, that a variant or cognate form of a particular word or discourse unit exists *until we know what it is* (and so, can search for it). We must, moreover, introduce caveats with respect to the medium being searched: such a parallel search exhausts an admittedly large, but also highly partial and non–random, sample of 'Early Modern Culture': only print, not manuscript or speech; only English, not Latin; only print that survives, not any of the lost pamphlets and ephemera, or books. Representative, certainly, but in a particular way.

And what *is* this corpus of texts? When we search it, will we be reclaiming cultural knowledge or creating it? After all, no one in the Early Modern period had access to all of these texts in this way. An early modern reader differs from a contemporary one precisely in the scale on which his or her indexical tools harvest data from the textual field (proper names, perhaps, but syntactical units: no). If, for example, we are researching the use of the word 'accommodated' in Early Modern print, we can be reasonably sure we have every surviving print use of this spelling of the word. We will therefore have read the word more often in its Early Modern contexts *than any actual Early Modern reader* ever did. What is the nature, then, of the knowledge we bring back from this search? If we are looking at the relationships between dramatic genres, we can include on our diagrams more plays than anyone in the period saw or read. Is this a problem for our claims about genre, if we think that 'genre' is a social process – a set of interpretations of texts, influenced by experiences of other texts?

Texts are artefacts: they have survived to be scanned, collated, data-mined and re-read. But we can point to other artefacts as well: the results of searches, the indexes that make such searches possible, and the statistical techniques applied to the results of these searches. When we ask, how many times does the word "thing" occur in Shakespeare's plays?, we are thinking in terms of concordances – in terms, that is, of a familiar technology for presenting variation within a text. The answer to such a question has been available since the first Shakespeare concordances occurred in the late eighteenth century.[19] But when, following this line of inquiry, we go one to say that the word "thing" occurs more frequently (above a threshold of statistical significance) in Shakespeare's works than in any other dramatist of his generation, we are saying something that no one in Tudor/Stuart London could have known as a statistical fact. To whom does this fact belong, to us or the early moderns? Put another way, is our knowledge of this word's *relative* pervasiveness a knowledge of something latent in the record of print? We can say that such relative frequency was potentially discoverable but practically "unknowable" during Shakespeare's lifetime. But if we insist that this statistical fact is and *was* nevertheless something real, we seem to have unearthed something latent, collective, structural or unconscious in the textual record of this culture. What else could such an abstract relation be if we are not calling it accidental noise?

Surely this distinction between knowable and unknowable, visible versus structural or unconscious, is itself relative: compare the "facts" we know today about the relationships among elements in the early modern print corpus to those we might know 100 years from now. Even when that corpus has been stabilized, the volume of discoverable future unknowns is potentially infinite. New units of discourse can be invented; new mathematical techniques for relating these units can be found. Where once it was the author's genius that guaranteed the authenticity of critically-discriminated patterns in Shakespeare's words, the level of semantic, syntactical and generic integration we are now contemplating in early modern studies is very clearly un-locatable in the consciousness of even the most gifted author. Like present-day scientific statements

about radiation that occurred before human life evolved in the universe to measure it, mathematical statements about once unobservable patterns in the textual record of the past introduce us to what philosopher Quentin Meillassoux has called "the problem of ancestrality." Such statements assert something real about the world (i.e., mathematically intelligible) which presumably would *still* be real if no one invented the means to observe it.[20]

The claims we make through parallel or distanced reading, algorithmic or iterative criticism, must therefore take the traditional hermeneutic question of "horizons of intelligibility" onto new, speculative terrain. It is a terrain that contemporary philosophy is already travelling.[21] Meillassoux introduces the problem of ancestral statements to undermine the linguistic determinism (and perhaps species-narcissism) of certain kinds of post-structuralist theory. We find this an apt provocation, one equally applicable to statistical discoveries made within a digitized corpus of texts. Certainly we will discover new things about Shakespeare, his contemporaries, and early modern culture as such in the post-transcription world of early modern studies. After the deluge, there will be criticism—but also, philology, statistics, linguistics, and old-fashioned literary history. What is perhaps most provocative about the findings we make in this post-deluge world, however, is their status as retroactive statistical facts – independent of subjects who could never have hoped to recognize them, but already real for us once belatedly discovered.

---

[1] See our initial work on the 'Very Large Diagram': Jonathan Hope and Michael Witmore, 'The hundredth psalm to the tune of "Green Sleeves": Digital Approaches to the Language of Genre', *Shakespeare Quarterly* vol, 61, no. 3 (Fall 2010), pp. 357–90.

[2] Franco Moretti, *Graphs, Maps, Trees: Abstract Models for Literary History* (London: Verso, 2005).

[3] As de Grazia notes, "If a lexical definition of a word requires a prescribed pronunciation, spelling, grammatical function, definition, and etymology, it may be possible to argue that *no* word in Shakespeare's time fully satisfied those conditions" (153), in Margreta de Grazia, 'Homonyms before and after lexical standardisation', *Deutsche Shakespeare–Gesellschaft West Jahrbuch* (1990): 143–56.

[4] See Jonathan Hope, *Shakespeare and Language: Reason, Eloquence and Artifice in the Renaissance* (London: Arden, 2010), ch. 3, 'Ideas about Language in Shakespeare,' 72–97.

[5] David Crystal, *Think on my Words: Exploring Shakespeare's Language* (Cambridge: Cambridge University Press, 2008), 6.

[6] Hugh Craig, forthcoming, 'Shakespeare's Vocabulary: Myth and Reality', *Shakespeare Quarterly*; and Ward E. Y. Elliott and Robert J. Valenza, 'Shakespeare's Vocabulary: Did It Dwarf All Others?', in Mireille Ravassat and Jonathan Culpepper (eds), *Stylistics and Shakespeare's Language: Transdisciplinary Approaches* (London: Continuum, 2011).

[7] Jürgen Schäfer, *Documentation in the O.E.D.: Shakespeare and Nashe as Test Cases* (Oxford: Clarendon Press, 1980).

[8] Terttu Nevalainen, 'Early Modern English lexis and semantics', in Roger Lass (ed.), *The Cambridge History of the English Language: Volume III 1476–1776* (Cambridge: Cambridge University Press, 1999), pp. 332–458.

[9] Marvin Spevack, 'Shakespeare's Language', in John F. Andrews (ed.), *William Shakespeare: His World, His Work, His Influence* (3 vols), (New York: Charles Scribner's Sons, 1985), vol. 2, pp. 343–61.

[11] Wordhoard is free and can be accessed at: http://wordhoard.northwestern.edu/userman/index.html. An explanation of log–likelihood can be found at: http://wordhoard.northwestern.edu/userman/analysis–comparewords.html#loglike.

[12] On exploratory analysis, see John W. Tukey, *Exploratory Data Analysis* (Reading, Mass: Addison-Wesley, 1977).

[13] See "'The Hundredth Psalm'" and our blog post on speech prefixes at http://winedarksea.org/?p=20.

[14] See our article, "The Very Large Textual Object: A Prosthetic Reading of Shakespeare," *Early Modern Literary Studies* 9.3 (January, 2004): 6.1-36. On our use of Principal Component Analysis, see our "Shakespeare by the Numbers: On the Linguistic Texture of the Late Plays," *Early Modern Tragicomedy,* edited by Subha Mukherji and Raphael Lyne (London: Boydell and Brewer, 2007).

[15] This work is ongoing, but we discuss the diagram in our *Shakespeare Quarterly* article and on the blog entries, http://winedarksea.org/?p=727 and http://winedarksea.org/?p=801. This result, although crude, shows us that *all* the sources of variation (author, generation, genre, editorial treatment, but perhaps also acting company, venue, print house) are expressed and so visible to statistical analysis. None can be deemed *a priori* decisive or "structurally" prior.

[16] Indeed, libraries could be understood as repositories of metadata which has yet to be applied to digital objects.

[17] See Daniel Shore, "WWJD? The Genealogy of a Syntactic Form," *Critical Inquiry* 37:1 (Autumn, 2010), 1-25. Shore writes: "A syntactic form's genealogical meaning emerges only when it is made simultaneously present alongside a plurality of other—and, if only ideally, all other—forms" (24-25).

[18] Shore argues, persuasively, that within the known texts available for indexical search, the counterfactual "would" with reference to Jesus first occurs in the Presbyterian divine Edward Reynolds' treatise, "The Life of Christ," in *Three Treatises of 'The Vanity of the Creature'…* (London, 1631), 427.

[19] Wells gives the date for the first Shakespeare concordance, produced by Andrew Becket, as 1787. See Stanley Wells*, The Oxford Companion to Shakespeare* (Oxford: Oxford University Press, 2001), 88.

[20] Meillassoux argues this point with respect to mathematically describable objects in the natural world, objects which he argues have an existence independent of (because indifferent to) the categories used by the Kantian subject. See Quentin Miellassoux, *After Finitude: An Essay on the Necessity of Contingency*, trans. Ray Brassier (New York: Continuum, 2008). An important test-case for Miellassoux's analysis of post-Kantian or linguistic "correlationism," would be that of statistical "facts" about *human* artefacts that could never have been known by those who created and used them.

[21] One can gesture, here, to the thriving school of "speculative realism" in which claims such as Meillassoux's are being debated. See, for example, the collected essays in *Collapse*, vol. II, "Speculative Realism," eds. Robin Mackay and Damian Veal (Oxford: Urbanomic, 2007).