

Data-Mining Shakespeare
A Folger Shakespeare Library Podcast

Michael Witmore
Director, Folger Shakespeare Library

Part of the Director's Choice lecture series

October 26, 2011

WITMORE: Hello, it's a pleasure to see you here tonight.

One of the great things about being the new Director of the Folger Shakespeare Library is that I've been able to meet so many new people and talk to a range of people as a scholar that I haven't really been able to address in my teaching. One of the funny things that also happens is that people now call me to ask what the library's opinion is about particular questions. So, in the last week I've had questions coming from journalists about the film *Anonymous* which is coming out this week. [LAUGH] And the question is what is the Folger Shakespeare Library's position [LAUGH] on this important controversy? And my answer first is libraries don't have positions, they have collections. [LAUGH]

If you're interested in the answer to this question and you think you can find the piece of paper on which the Earl of Oxford wrote, "I, the Earl of Oxford wrote all of Shakespeare's plays and with the following coconspirators managed to pull off this incredible fraud," [LAUGH] you'll probably find that piece of paper if it exists in this library. [LAUGH] So get a reader's card and come on in. But I'm confident that that piece will never be found and that we will never have to change our name to the Folger Oxford Library. [LAUGH]

But I'm happy not to speak as Director tonight but as a scholar and a researcher. I have come to digital studies in a roundabout way and I'd like to talk to you a little bit about that tonight. I'd also like to open up a series of questions that I think are important for libraries, scholars, and well-read, interested people about the nature of literary history.

So let's begin. Our topic tonight is data-mining Shakespeare. The title is deliberately provocative. Data-mining is something that we do in epidemiology. It's something we do in the social sciences but it is not something that we associate with literary studies. In fact, I'll address this question in a moment.

I myself as a scholar was trained to read books slowly and if there's a literary equivalent of whatever the slow food movement is [LAUGH] I think that's the right way to read books. And when scholars are in the reading room that is what they are doing. They're going word by word, line by line, trying to understand a historical context, a person, a writer, or a problem.

So having been trained in that way of thinking and researching I was surprised to find that there were interesting things to be said and learned from reading quickly or not reading at all. [LAUGH]

And that's what I'd like to talk about tonight. First, my own sense of shock that one would actually want to do something besides sit and read but also to talk about what it is that you could learn by not reading. So this is the research question that I have and I share it with a number of colleagues in the UK, in the United States, Canada. We're interested in variation. What does linguistic variation, and by that we mean the use or omission of words or phrases in texts, tell us about history, culture, and interpretation?

I'll start with the tool that we use to do this work. My first job as an academic was at Carnegie Mellon University where I was teaching in the department of English. While I was doing my slow reading and teaching classes in Shakespeare, I met a man named David Kaufer who happened to be the chair of my department. I was a junior faculty member trying to write my first book, trying to get tenure, and Professor Kaufer would come into my office once every other week and say "I've got this terrific contraption that reads books without reading them, you've got to try it." [LAUGH] And my first response was no, I don't. [LAUGH]

So after several months of this I realized that I needed to do something and I said, "I'll tell you what, I will give you 36 plays written by Shakespeare and these 36 plays were published in the First Folio in 1623. Shakespeare's friends made a table of contents for this book and divided those plays into three groups. I will give you text files for those plays and I'd like you to sort them into groups." And I figured that that would be the end of it. [LAUGH]

Five minutes later he was in my office [LAUGH] with a diagram that perfectly took all of Shakespeare's history plays, put them in a group, but it took out *Henry the VIII* and put it next to the late plays because *Henry the VIII* is really a late play or a romance. It took the comedies and put them in a group and it spread the tragedies around a bit.

And my first thought was how in the world could this man have done this? The device that he created called DocuScope was a device created for teaching freshmen writing.

What it does is take a text which you have fed into the machine. [It tags the different words in the text and then it shows you by color coding those words what kinds of words you are using in your writing.

Now it turns out this is not really a great tool for teaching freshman writing [LAUGH] and that's an interesting thing in and of itself. Why is this not an especially useful tool? Because for native speakers of English we often do many things unconsciously that we simply do because we know the language. And having your attention called to those things may not necessarily be the most productive way to learn to write. But if English is your second language, having your attention called to certain types of words can be very productive. And so this tool which is now no longer used to teach freshman English is used to teach writing in Qatar and in China. So the tool is doing something else.

It's also, inadvertently, I think, from the creator's standpoint, being used to study literary texts in English print.

So our adventure with this tool began with a colleague of mine in the UK named Jonathan Hope who is a historical linguist, began when we took a tool that was intended for one thing and started using it for another.

Our first experiment we called "Shakespeare in Pieces."

This is the catalog page of the first folio, a remarkable book of which the Folger possesses 82 copies. There is one in the lobby.

Hemmings and Condell have divided the plays up into three groups, comedies, histories, and tragedies. As eye witnesses to the early modern theatrical scene we can use their judgments to talk about what had to be in place for someone who knew these works to call them that, that, and that.

So my colleague in computer science would say you have domain scientists from the seventeenth-century who understand the phenomenon who made a judgment and we are going to reconstruct the criteria that they used in order to make that judgment but we're going to do it by different means.

I suspect that these judgments about why comedies are different from histories and tragedies were made on the basis of plot and what we know of the reception of poetic theory and the Renaissance tells us that plot was an important way of understanding what kind of literary work you were looking at.

I'll just give you an example: when I teach my Shakespeare class—the survey of Shakespeare's works—we start out doing some comedies. My students say comedy is funny. And I point out that in the Renaissance the word comedy usually referred to a play that ends in a wedding. It also often refers to a play that will have characters from multiple social strata.

Those I think were the criteria that were being used in the Renaissance and they're also a criteria that we use. We are very good at following plot. We're also extremely good at following tone. If I ask you if you've gone and seen our production of *Othello*—which is a wonderful production, “hat is a tragedy?” I think you would say a tragedy is a play that ends badly.

Now we could say a lot more about that but the way we describe that is in terms of the ending. What if there are other ways of capturing what makes comedy, history, and tragedy themselves, not something that you and I would use when we talk about plot, but something that's more from a worms-eye view, something that only a computer could keep track of on the level of a sentence? And so this is our discovery. Things that we recognize as genres of writing and literature are visible on the level of the sentence.

That's interesting because what it means is that there's a kind of integrity to our language which goes all the way down. Shakespeare to tell a story in a room like this with twelve actors in two hours without a soundtrack, without flashbacks, was going to have to do certain things. It's a product in part of the nature of theatrical performance. And it's those things that we're picking up at the level of a sentence.

So how can we describe these divisions in a way that does not start with plot. This is the first alienating diagram [LAUGH] that I'm going to show you, if you are a slow reader.

In the course of trying to figure out exactly how Kaufer had managed to produce these groups I learned about something called principal component analysis. It is one of the techniques that's used in data mining. Also factor analysis is used and there are even fancier kinds of procedures which we have tried but which are maybe too fancy. This is something that most statisticians understand. It's a well known procedure.

Let me explain how it works and what this is. First, these are 767 pieces of Shakespeare's plays in something called PCA space. What are the pieces? Pieces are one thousand word segments of plays. Why are we working with tiny pieces? Because to do PCA you need to have lots of little items if you want to take advantage of counting lots of different kinds of things. The statistician will tell you that you need to have more instances or trials than you have variables. What this means is that given what DocuScope counts, we need to look at smaller and smaller pieces. So that's from statistics.

Now principal component analysis is looking at each one of these pieces and saying there are things here that I can count. We start with DocuScope. The contraption DocuScope can count somewhere around 200 million strings of English and place them into 101 buckets. These 101 buckets were created by David Kaufer over the course of ten years by going through the Oxford English dictionary first and taking words and putting them in the buckets. That's eight hours a day for ten years, eight million discreet acts of coding.

Once you count things in a text with DocuScope you then have a spreadsheet or a table which says my first item has this much of bucket number one, this much of bucket number two, this much of bucket number three, and you get percentages all the way across.

What are those buckets? [LAUGH] Those buckets evolved over the course of Kaufer's work to capture what he thought of as significant types of language use. So if you want to accomplish a certain task as a writer you have to do these 101 things. Is that number arbitrary? Yes. Is this structure subjective and also, therefore, arbitrary? Yes.

What DocuScope produces is a highly systematic and subjective view of our language. It is a kind of private rhetoric or theory of language that is made operational with the help of a computer. I do not think a device like this will be built again in the next few decades. [LAUGH] But I was lucky that I was in a department with a man who had done it and thought that it would be nice to try it on something else.

So, you take the spreadsheet that has the measurements of every fragment of Shakespeare's plays according to these 101 functions or variables. And then a computer can go through and say, well, after having looked at the 767 pieces, I notice that whenever an item has variables a, b, and c, it doesn't have d, e, and f.

That's something that you and I could never keep track of. Imagine taking 36 decks of cards that are filled with random contents and looking at them for a few days and then when someone asks you, "You know, whenever I saw a deck that was full of red cards, I had very few sevens." That's the kind of massively comparative work that is happening in PCA.

For humanists, it takes a while for us to understand what that is and why you would be interested in it. But once you realize that a principal component is a recipe for having and not having certain things, it becomes a lot more interesting.

The first principal component in an analysis is by definition the most pervasive pattern in the collection. I'm just going to get a drink of water here. We never had water when I was lecturing at the university. [LAUGH] So it's the pattern that's visible from the farthest away. If it turns out that all the decks that have red cards never ever, ever had a seven that would be the first principal component. And then it goes down to smaller and smaller patterns that it's seeing.

We then take those components and graph them on an x and y axis. So we're then rating every item on how it scores according to those visible patterns.

When we took the first principal component and the fourth principal component and put them together we found an interesting result. That is, items in the lower left corner that are low on principal component four and low on principal component one tends to be history plays. Those green dots that are in the lower left are Shakespeare's history plays.

Second thing that is interesting about this diagram: there are not a lot of red dots there, the red dots are up here. This is the single most important fact I think linguistically about Shakespeare's writing. The things that he does when he is writing comedy are the things that he does not do when he's writing history. We could call them anti-types. That difference is visible from far, far away.

And in particular, Shakespeare's history plays as a genre and a kind of writing holds together very tightly based on what they do and don't do with words. I think that's because Shakespeare started with history plays. It's a genre that he helped perfect and he did the same thing over and over again.

This is a worms-eye view of what he's doing. It's not the description that you or I would give. The types of designations we have here, this is what we call metadata genre: blue, late plays; brown, tragedy; red, comedy; and green, history.

So here's our diagram—well, what's in it? On the upper right, there is an item that is extremely high on principal component one and principal component four. If you really wanted to be hyperbolic you would say this is the most comic thing Shakespeare ever wrote. [LAUGH] It's a passage from *The Merry Wives of Windsor*, 2.1. I've given this to you on your handout and I've given it to you as a text which is what you and I read and I've also given it to you as a screenshot from DocuScope. I'll talk about this but you have the other one to look at.

What DocuScope is doing is underlining the words that it has classified as being important for pushing one of these play chunks into that upper right-hand corner. So we ask ourselves why are those things in the upper right-hand corner? The answer has something to do with the words that are underlined here. This is the moment in humanities research where we get to, as a group, look at the passage and start asking why. It's potentially, I think, the most exciting consequence of this kind of work because we rarely get together in groups, look at patterns, and say I think it's this. I think it's this, I think it's this.

One of the things we can do with DocuScope is ask why have you colored all these words in red? Those are a part of a DocuScope category called first person and we have the percentage of the entire text that is first person. It is: I, mine, I. There's also another type of string or sequence called self disclosure and "I had rather" is part of it. It's usually the first person pronoun with a verb. Often it's a tensed 'ed' verb: I walked, I saw, I kicked. That's what's red.

What we get in orange is something called—and these were names given by the program's creator—uncertainty. So words like indefinite pronouns: some, no not, doubt, mystery. There may be times when the word mystery is not expressing anything uncertain. The philosophy behind DocuScope is you pick the function you think is most likely and you go with it and you count it that way every single time. So when we find things that I disagree with and say, "I don't think that word serves that function, I take comfort from the fact that it's going to be wrong the same way every darn time [LAUGH] and I could correct for it." F

inally we have something called direct address which is you, let's, find you, thy, thy, you, ever. Why is this passage exemplary for what comedy does and doesn't do? Comedy often has transactions between two people, not more, who are talking about a plan, something that is going to take place offstage, who are comparing opinions and weighing them, or, and I think this is very common, who are in a state of what you could

call congruent misunderstanding. [LAUGH] So they kind of understand each other but they don't understand each other and they don't resolve it. It just keeps going.

Now at the end of a comedy this I, you, I, you can shift. It can become we. Given what we know about the plot of comedy this finding makes perfect sense. One way of describing comedy is it is the genre of writing that doesn't allow you to say we until act five. [LAUGH] Why? Because there is a marriage.

I'll come back to this pattern in a moment but I want to show you something that is its opposite. Because these two types of writing cluster at opposite ends of this spectrum, we can have the least comic piece of writing which is down here. This is a passage from *Richard II*. It is a heraldic moment where Bolingbroke and Mowbray have come to throw their gauges down and to accuse each other.

Now what is it that has to happen here that is not happening in comedy? You can see that a lot of this passage is underlined. What are those words? One of them is something called descriptive features and these are the items that are in yellow. Words that describe sensible properties of things, objects which can be sensed themselves, spatial relations among objects, and motion.

Why is it that history would have that kind of word in it? Well, if you want to tell a story about battles and armies and personalities—and as the prologue says in *Henry V*, you cannot bring them all onstage—you have to do it by reporting. History is the genre that talks about things. Not the usual way you would think of history but from a linguistic standpoint, it's true enough.

What does it also do? It uses words that refer to a common authority. I think this is a very clever thing to count. What Kaufer did was say, "If you're looking at a situation where people have to name check the kind of authority that is supposed to govern multiple people, you're in a context where people are assuming social consensus and that institutions exist which have some purchase on your deliberation and your moral thinking. If you want to depict that kind of society you'll have to use these kinds of words once in a while." And history does it often.

Finally, history uses these words in purple. These are known as person property. Here is a tagging that I would disagree with. I don't think Henry is a person property. It's a proper name, but king, combatants, warder, infant, cousin—these are all social roles. The only other place where I see this in comedy is in something called city comedy

which is a subgenre of comedy that develops which focuses on the professions and types of social functions of individuals in urban London.

So, between these two passages that I've shown I've tried to show you what history does, what comedy does, with the knowledge that history lacks those things that you saw on the previous slide. It's always a combination of having something and not having something that produces the tight grouping. And that's something that's very hard for us, even for really slow readers to get. I can pay attention to every single word that comes by but I have a hard time paying attention to what's not there. That is why data mining and PCA are useful for thinking about genre. Because genre is a coordinated set of having things and not having things at least from a linguistic point of view and you could not see that.

The red dots are comedies, the brown dots are tragedies. So, what's that? It is *Othello*.

About 30 years ago a reader who worked in the Folger Library, a scholar named Susan Snyder said, "You know, there's something very interesting about *Othello*. It's written like a comedy." And after her, several other Shakespeareans started to look at this play and say, "You know, many of its generic components seem to have come from comedy. What's different is that last twist at the end is devastating."

If you saw the performance on Monday night, one of the things that I thought was extremely interesting about the production—and if you come to see the play I think you'll really notice this—is that many of the moments in the action onstage are played as comic. Essentially what's happening is Ian Merrill Peakes is addressing someone onstage doing the "I / You" game that we identified, not talking in a rich way about objects, motions, things like that. And from the side he's letting the audience know that what I'm saying to this "you" is not what I believe.

How did Shakespeare accomplish that particular thing onstage? He had the actor who's playing Iago speak to two different people at once. I think what's really happening there—it's something that happens also in Shakespeare's *Twelfth Night*—you get two characters who are talking to each other about someone else. They're plotting. You get lots of plotting in comedies, whether it's trying to get Benedick and Beatrice together, whether it's trying to get Falstaff to tip his hand so that you can dump him in the mud and make fun of him. What's happening in *Othello* is that this other person that's being discussed is probably either the audience itself or the true Iago and that Shakespeare figured out a way to make this pattern work and to push out a kind of irony for the audience that is built into this structure.

The other scene that looks a lot like this is the scene from *Twelfth Night* where, if you remember this play, Viola, who's been shipwrecked, dresses up as Cesario. And then there's this moment when she and Olivia are speaking to each other and Olivia really wants to talk to Cesario because Olivia loves Cesario but Cesario is a she, not a he. And I think what Cesario is doing in this back and forth is being a third person in this dialogue. There's the real Viola, there's Cesario who's doing the talking, and there's Olivia who's being addressed. So, Shakespeare figures out ways of instead of keeping this third person offstage when you're plotting of actually making one of the two people talking be that third person and that is producing specific theatrical effects. I think it's something that he does in a very deliberate and devastating way in *Othello*.

What makes *Othello* a comedy in part is the fact that the two male characters are plotting against a third, Othello's wife, and that the "we" that emerges at the end of the comedy is the pairing of Othello and Iago. It is a romance that excludes the beloved. I think that is absolutely deliberate and I think it's one of the things that makes *Othello* a truly awful play [LAUGH]. It is perverse in the way that it takes the kinds of things you do when you're writing comedy and sets you up emotionally to expect one thing and then turns on you. That is what makes that play so awful. Another way of describing that is saying *Othello* is a play about a couple who love each other, a terrible mistake in judgment, and a death. That's also a description of what's happening.

Do I prefer one over the other? Well, if I really want to know what *Othello* is I think I'll just run it through DocuScope and figure out what it said. [LAUGH] Or maybe I should go talk to Susan Snyder. Well, I should do both. They're both ways of getting at something that is incredibly complex.

With Shakespeare we have a theatrical mind that thinks in this myriad way about what he can do with words and then what kind of plot he can build on those words. And the result of this kind of work is that you start to think of plays and their words as part of a built environment. What do you do with a play? It's a contraption; you build it and you perform it in a space like this.

These are photographs taken by Muybridge at Stanford University. Muybridge was interested in animal locomotion. By staging a series of cameras on the Stanford campus and having a horse and rider ride across the frame, each step in the sequence tripped off a camera that took a shot, he was able to simulate what we would now do with film. What's interesting about this is that Muybridge was able to see something that you can't

see with the human eye. He saw that a horse in a full gallop lifts all four feet off the ground. I'd like to close with this thought and then we can talk if you have questions.

We as human beings are really good at thinking about plays, texts, and writing from the waist up; that is, we look at the hand gestures, the expression on the face. But we don't pay attention to what's happening from the waist down. Shakespeare from the waist down is really interesting [LAUGH] because the things he does with words, the footsteps, right?—Where do you have to put your feet? What kinds of pronouns must you use? What kinds of verbs are you going to use so that you can with your upper body do all of these different things, comic, tragic? They're connected and there's a way in which the things that we can count really are the weight-bearing members of texts and plays.

What is the more real thing? Is it what's happening with the writer on top or the fancy footwork below? They're both equally real. But it depends on what you're interested in and what you're trying to learn.

In the long run, I think that this type of work is only going to raise more and more interesting questions. We have 27,000 texts from the early modern period digitally transcribed and my colleagues and I have begun to work on the next largest order of magnitude that we can handle, about a thousand texts written over the course of 270 years. Once you start getting big number of texts it's hard to interpret the results because there is no way you could have read all of that. So you need a group. A research library with people who have actually read the material is probably the only place in the world where these diagrams can be interpreted. Otherwise, you might as well be studying weather patterns at the equator because the graphs will look the same.

The real struggle here I think in the long term will be to take what we know in the rich comparative way that we know it about a few texts and help it lead us to things that we don't know and encourage us to read things that we never would have thought to read.

I hope that in the next ten years we have transcriptions of everything that was printed in English from 1470 to 1700. I think it will happen.

What would be an interesting question to ask of that magnitude of a collection of texts? Well, I think the first would be: what are the distinctive types of European writing that emerge out of the scientific revolution, the coalescence of nation states, and mercantile capitalism? Why is that the novel appears somewhere in the middle of the seventeenth century and where on earth did it come from? Did it come from newspaper reporting,

Folger SHAKESPEARE LIBRARY

Advancing knowledge & the arts

histories, travel logs, scientific writing—all candidates for the genetic precursors of the novel? Maybe we can learn something new about that and if we can tell that story, the next question is once printing becomes a fixture of the entire Atlantic, North America, potentially the Caribbean, texts written in Africa, at what point do we begin to see texts in North America that more resemble North American texts than they resemble anything written in Europe? I think we will know the answer to that question in ten years. And I suspect that that the things that we come up with to answer that question, we will be surprised. We will find that the first most American text written or printed in English was a sermon, an agriculture manual, an almanac. There's no way to know.

Thank you for listening. I look forward to your questions and I'd be happy to answer them. [APPLAUSE]

