Open Library of Humanities

Part of the Ubiquity Partner Network

Digital Studies / Le champ numérique

Research

How to Cite: Estill, Laura and Luis Meneses. 2018. "Is Falstaff Falstaff? Is Prince Hal Henry V?: Topic Modeling Shakespeare's Plays." *Digital Studies/Le champ numérique* 8(1): 1, pp. 1–22, DOI: https://doi.org/10.16995/dscn.295

Published: 23 January 2018

Peer Review:

This is a peer-reviewed article in *Digital Studies/Le champ numérique*, a journal published by the Open Library of Humanities.

Copyright:

Open Access:

Digital Studies/Le champ numérique is a peer-reviewed open access journal.

Digital Preservation:

The Open Library of Humanities and all its journals are digitally preserved in the CLOCKSS scholarly archive service.

Digital Studies / Le champ numérique

Estill, Laura and Luis Meneses. 2018. "Is Falstaff Falstaff? Is Prince Hal Henry V?: Topic Modeling Shakespeare's Plays." *Digital Studies/Le champ numérique* 8(1): 1, pp. 1–22, DOI: https://doi.org/10.16995/dscn.295

RESEARCH

Is Falstaff Falstaff? Is Prince Hal Henry V?: Topic Modeling Shakespeare's Plays

Laura Estill[1] and Luis Meneses[2]

[1] Texas A&M University (TAMU), US

[2] University of Victoria, CA

Corresponding author: Laura Estill (lestill@tamu.edu)

This essay demonstrates how topic modeling can be fruitfully applied to TEI-encoded plays, which allows scholars to analyze speeches by individual characters. Our analysis centers on Shakespeare's corpus and characters who reappear in multiple plays. Specifically, we use topic models to show that young Prince Hal (in *1 and 2 Henry IV*) does not speak the same language as his later self, Henry V (in his titular play): his linguistic shift mirrors his shift in status. Hal himself announces, "I have turned away my former self"—his change in diction bears out his assertion. Conversely, topic models reveal that Falstaff is Falstaff across multiple plays and genres (notably, *1 and 2 Henry IV* and *The Merry Wives of Windsor*), despite scholarly claims to that the Falstaff of comedy is a watered-down version of the braggart drunk of the history plays. Ultimately, we hope that this algorithmically-informed analysis of Shakespeare's plays is not taken as a final answer, but, instead, as a prompt. As this research reveals, topic modeling plays with attention to each speaker opens the door for new comparisons, and in turn, expands on previous interpretations of literature.

Keywords: Shakespeare; drama; topic modeling; Falstaff; Prince Hal; *Henry V*; *Henry IV*; *Merry Wives of Windsor*

Cet essai démontre que les modèles à thèmes (*topic model*) peuvent être appliqués avec succès à des pièces encodées en TEI, ce qui permet aux érudits d'analyser le discours de personnages individuels. Notre analyse se concentre sur le corpus de Shakespeare et sur ses personnages qui réapparaissent dans plusieurs pièces. Particulièrement, nous employons des modèles à thèmes pour montrer que le jeune Prince Hal (*1 et 2 Henri IV*) ne parle pas le même langage que celui qu'il parle après être devenu Henri V (*Henri V*): son changement linguistique reflète son changement de standing. Hal, lui-même, annonce: « j'ai renoncé à mon passé » —son changement de diction confirme cette affirmation. Inversement, les modèles à thèmes révèlent que Falstaff est Falstaff à travers plusieurs pièces et

genres (notamment, *1 et 2 Henri IV* et *Les Joyeuses Commères de Windsor*), malgré des affirmations érudites que le Falstaff dans la comédie est une version édulcorée du vantard ivre des pièces d'histoire. Finalement, nous espérons que cette analyse algorithmique des pièces de Shakespeare n'est pas considérée comme une solution finale, mais plutôt comme une réplique. Comme cette recherche le montre, l'usage de modèles à thèmes pour analyser des pièces, ce qui se concentre sur chaque personnage, offre de nouvelles voies de comparaisons et étoffe donc nos interprétations de la littérature.

Mots-clés: Shakespeare; pièces de théâtre; modèle thématique; modèles à thèmes; Falstaff; Prince Hal; *Henry V*; *Henri IV*; *Joyeuses commères de Windsor*

Taking into account Shakespeare's pre-eminent position in the English literary canon, there has been surprisingly little work on topic modeling his corpus.[1] The relative lack of topic modeling on Shakespeare could be attributed to the fact that his most popular works today are plays; rather than having an overarching narrative or authorial voice, Shakespeare dramatizes many characters with distinct voices. For this study, we apply topic modeling algorithms to XML-encoded plays (that differentiate between speakers) in order to analyze the relationships and the recurring themes between the characters in Shakespeare's plays.

One of the benefits of topic modeling is that it can lead researchers to new avenues for comparison. Keyword and thematic analysis has always been a mainstay of humanities criticism; as this essay shows, topic modeling offers a mode of literary analysis that expands our understanding of word use beyond thematics and pre-selected keywords. With the appearance of Martin Spevack's *Harvard Concordance to Shakespeare* in 1974, scholars could easily find and compare Shakespeare's use of particular words across plays (Spevack 1974). The advent of digital text, such as MIT's *Complete Works of Shakespeare* (Hylton 1993) and *Open Source Shakespeare*

[1] There has been non-linguistic and hierarchical topic modeling applied to Shakespeare's plays—see, for example, Duhaime 2014 and Schaefer 2015. Sentiment analysis, sometimes considered a form of topic modeling, has also been applied to Shakespeare's plays, see, for instance, Nalisnick and Baird 2013a and 2013b. In her 2017 Shakespeare Association of America paper, Amanda Henrichs applied topic modeling using R to Shakespeare's sonnets, but not the corpus of plays.

(Johnson 2003), made Shakespeare's works easily searchable with a command-F or control-F. Yet this form of inquiry assumes that scholars know what terms or topics are worth searching. Topic modeling makes no such presumptions; it provides scholars with possible terms and topics to pursue further. Topic modeling, of course, is not analysis unto itself, but it can promote further analysis of literature and in the humanities.

The purpose of our study is two-fold: first, we apply topic modeling to identify patterns and hidden trends in Shakespeare's plays; and second, we analyze these patterns and justify them using the conclusions gathered from previous scholarly work. While topic modeling can point us to patterns within and across Shakespeare's plays, accurately characterizing the nature of these relationships is the real challenge. Specifically, we use topic models to show that young Prince Hal does not speak the same language as his later self, Henry V: his linguistic shift mirrors his shift in status. Conversely, topic models reveal that Falstaff is Falstaff across multiple plays and genres.

Background

Topic modeling is a concept that is often used to describe a set of algorithms that are used to describe a large number of documents that often share a common theme (see Blei 2012). In other words. topic modeling algorithms are statistical methods that analyze the words in texts to discover the themes that run through them, how those themes are connected to each other, and how they change over time. More importantly, topic modeling as a technique makes no previous assumptions about the text and does not require any prior annotations or labeling of the documents —which is one of the reasons why we favored its use. As a consequence, the topics themselves emerge from the analysis of the original texts. Topic modeling enables us (as researchers) to organize and summarize documents at a scale that would be impossible by human annotation —or that would require large efforts to do so.

Latent Dirichlet Allocation (commonly known as LDA) is the simplest form of topic modeling (see Blei and Jordan 2003). The main idea behind LDA is that the text in documents do not belong to a single topic exclusively, but can belong to multiple topics at the same time instead. Formally, a topic can be defined as a probability

distribution over a fixed vocabulary. For example, a topic about English literature has words about literary works with high probability, whereas a topic about computer science has words about engineering with a high probability.

LDA can be explained using two steps. First, we randomly choose a distribution of words over topics. And second, for each word in the document we randomly choose a topic from the distribution over topics in step #1. As part of this second step, we also randomly choose a word from the corresponding distribution over the vocabulary. This process is usually carried out in multiple iterations. Therefore, a topic model with greater iterations is usually considered to provide a better representation of the data (see "methodology" below where we explain choices specific to this analysis). Consequently, the output of an LDA model usually consists of the probability distribution of documents and terms over different topics.

In our case, we specifically chose to use LDA because of its simplicity and, more importantly, because it makes no previous assumptions over the contents of the documents. Our intention was to fully take advantage of the affordances that topic modeling provides as a machine learning technique and discover the relationships within a collection of documents, which falls right under the scope of LDA. Since we are dealing with a collection of Shakespeare's plays these relations translate to the hidden connections between characters expressed in their dialogue.

Lisa M. Rhody (2012) has shown how LDA topic models can fail for figurative language. Indeed, despite the popularity of topic modeling and LDA models for critical inquiry, the "black box" nature of topic models and algorithms warrants attention: as Andrew Goldstone and Ted Underwood (2014) explain of their topic models, "No matter how carefully prepared the topic model may be, it is still important to remember that is a model, a statistical simplification." These caveats hold true for the research presented here. Rather than viewing a topic model as a critical endpoint, we position the topic model as a means to a critical end: the models do not offer the final analysis, but rather, they provide a lens through which we can approach the text. Benjamin J. Schmidt (2012) compellingly argues that "humanists need to ground the analysis of topic models in the words they are built from" (see

also Schmidt 2016). By using a relatively small corpus (Shakespeare's plays) and paying specific attention to characters and the words they speak, we follow Schmidt's directive. Indeed, this research embodies the spirit of digital humanities because it uses a digital tool as a way of sparking humanist engagement with the texts.

Methodology

Our research question—"(how) do Shakespeare's characters speak differently from each other?"—arose from our work putting together Digital Acting Parts (Estill and Meneses 2014), a website that allows actors and students to learn lines from Shakespeare's plays by mimicking and enhancing the early modern acting experience—that is, learning your lines from a 'part' instead of a full play. In Shakespeare's day, actors did not use a full script to rehearse: rather, they had an acting part, which was an often hastily handwritten manuscript with only their cue lines and their lines (Palfrey and Stern 2007). Digital Acting Parts is built on the XML files, schema, and text provided by the Open Source Shakespeare. For our topic modeling, we also drew on plays from Open Source Shakespeare, which have modernized spelling.

As Julia Flanders (2012) and Rebecca Niles and Michael Poston (2016) have argued, encoding is editorial and interpretive. In this case, we found that working with the already-encoded texts from Open Source Shakespeare for Digital Acting Parts raised a seemingly simple question: if we wanted people to be able to learn the lines of one character, how do we define a Shakespearean character? While the answer might seem simple (we could, for instance, turn to M. H. Abrams's definition that "**Characters** are the persons represented in a dramatic or narrative work" (1999, 32), Shakespeare's plays challenge overly pat responses by presenting some of the same characters in multiple works. Particularly in Shakespeare's two tetralogies, the historical figures Shakespeare dramatizes appear in more than one play. While for Digital Acting Parts, we decided to consider each character's appearance in different plays separately (asking users to first choose their play then their character), the question remained: is a character who appears in more than one play ultimately the same character? Or, to put the question more succinctly: is the Prince Hal of *1 Henry*

IV the same as Prince Hal in *2 Henry IV* and the same as the titular king in *Henry V*? This question has an important bearing on how we choose to analyze and stage Shakespeare's plays, whether as discrete units or as broader arcs that extend beyond a single book or performance.

To test the hypothesis that Shakespeare's recurring characters are the same across plays, we decided to focus on two characters from the Henriad, Shakespeare's second tetralogy: Prince Hal and Falstaff. There is no scholarly disagreement about, for instance, whether Henry IV is the same character in *1 & 2 Henry IV*: he clearly is. The changing or different characters of Prince Hal and Falstaff, however, have been scrutinized for different reasons. On the one hand, Prince Hal (who appears in *1 & 2 Henry IV and Henry V*) undergoes a major character transformation over the course of these three plays. Falstaff, on the other hand, bridges genre: he appears in *1 & 2 Henry IV* (histories) and *The Merry Wives of Windsor* (a comedy).

In order to apply topic modeling algorithms to our dataset of Shakespearean plays, we employed special parsers to transform the data. In the end, we parsed lines from 37 plays: all plays commonly attributed to Shakespeare, except *Two Noble Kinsmen*, which is not in Open Source Shakespeare. The plays from Open Source Shakespeare were originally in Microsoft Access format, which is a proprietary standard that we transformed into a TEI-compliant XML using two open-source Python libraries (Rossum 1995): PyMySQL[2] and lxml.[3] Then we extracted the lines for each character in every play using the same parser that we are currently using in Digital Acting Parts. The lines for each speech were then tokenized, lower cased, stemmed with a Porter stemmer to improve the accuracy of the results. We also removed stop words from the dataset, which we filtered using Python's Natural Language Toolkit (NLTK) corpus of English stop words. Additionally, we also removed terms that had fewer than three characters from the dataset. Then we used Gensim (Řehůřek and Sojka 2010), a Python library used for topic modeling, and its implementation of

[2] PyMySQL: Pure Python MySQL Client. (9/28/2016) Available: https://github.com/PyMySQL/PyMySQL.

[3] lxml Project. (9/28/2016). lxml – XML and HTML with Python. Available: http://lxml.de.

the Latent Dirchlet Allocation (LDA) model to cluster the speeches of the characters in the plays and obtain our results. Our LDA model is a hierarchical probabilistic generative model that can be used to model a collection of documents by topic (Blei, Ng, and Jordan 2003).

We performed several runs of the LDA model varying the number of topics and the number of passes. Taking into account the research methods found in previous work (notably Schaefer 2015), we finally settled on 50 topics and 100 passes, as it provided an appropriate clustering of the speeches of the characters in the plays. Specifically, the number of topics indicate the number of clusters for the documents while the number of passes increases the accuracy of the model. (After 100 passes, undertaking further passes was not yielding significant differences). The appropriate value for these parameters is dictated by a compromise between the computing time needed to cluster the documents and the validity of the results. The topics we developed included stage directions. As Appendix A shows, however, the stage directions clustered in two main topics with exceptionally high probabilities. In the original XML files that we used for Digital Acting Parts, lines were encoded for each particular character. Thus, this encoding treated the stage directions as a "silent" character in the plays. The fact that the stage directions were identified across plays and genres shows the effectiveness of our topic model. Furthermore, the words in our topic models are sometimes truncated, which is a consequence of the stemming that we did in the preprocessing stage. Examples of this include abbreviations such as "franci" and "princ," which can be found in the probabilistic term distribution that we extracted from our model.

Our approach to topic modeling Shakespeare's plays is innovative because it combines topic modeling with markup to analyze the parts of the text and not just consider texts as monolithic wholes. Mikhail Bakhtin famously explicated the "polyphony" of literary novels: that is, the differing voices and viewpoints expressed by characters within one literary work (1981). As Marvin Carlson points out, theatre offers "even richer possibilities" for Bakhtinian polyphony, as it brings together different speaking characters as well as actor, director, designers, rather than simply

having the words mediated by a reader (1992). Topic modeling with attention to *who* is saying *what* is one of the first steps towards recognizing the inherently polyphonic nature of theatre, and we hope this study encourages further work in the field.

Turning to Shakespeare

Our analysis shows that topic modeling is able to highlight the hidden relationships between characters in different plays. More specifically, the topic modeling analysis that we conducted sheds some insights about how characters evolve across different plays. For instance, topic modeling Prince Hal's speeches show how his character changes over the course of *1 Henry IV*, *2 Henry IV*, and *Henry V* by analyzing the words he speaks. Shakespeare famously included one of his most popular characters, Falstaff, in *The Merry Wives of Windsor* (a comedy) after he was a hit in the *Henry IV* history plays, leading centuries of scholars to decry "the Falstaff of *Merry Wives*" as "a big-bellied imposter" (Coleridge 1851) who is only a pale imitator. As Phyllis Rackin and Evelyn Gajowski put it, "By the beginning of the twentieth century the authenticity of Falstaff in *Merry Wives* had been fully discredited" (2015, 6)—he was even called "pseudo-Falstaff" by Harold Bloom (1998), as Rackin and Gajowski point out. Topic modeling Falstaff's language shows that Falstaff of the comedy is *not* an imposter: rather, to the statistical model, he is indistinguishable from Falstaff of the histories. Furthermore, though our topic modeling analysis we were able to find unexpected relationships between characters: for instance, King Henry V and Petruchio (from *Taming of the Shrew*) draw on similar lexicon even though they appear in plays of different genres, that is, history play and comedy, respectively. Future work in this area could focus on unexpected parallels between characters: topic modeling Shakespeare's plays by character leads to hitherto unexplored sets of literary foils, that is, perhaps unexpected pairs of characters to compare.

Some of the topics that our model extracted showed the thematic unity of individual plays. For instance, some clusters appeared focused on the characters of one play, such as 'Angelo,' 'Claudio,' 'Duke,' 'prison,' 'brother,' and 'good,' a topic primarily taken from *Measure for Measure*; the topic 'Hamlet,' 'king,' 'Horatio,' 'Laertes,' 'Rosencrantz,' 'Guildenstern,' 'queen,' 'Polonius,' 'Ophelia,' 'ghost,' and

'Denmark' clearly refers to the play *Hamlet* and actually offers little avenue for exploration as it only tells us that these characters often interact and belong in a group, which, of course, we already know from our dramatis personae.[4] Others, however, showed how some topics extend beyond a particular play. One of our topics, for instance, had 'Macbeth,' 'Banquo,' 'hail,' 'sir,' 'king,' 'murder,' 'witch,' and 'Caesar' as high ranking terms, which suggests that the model found the similarities between *Macbeth* and *Julius Caesar*—not least of which, of course, is prophecy by witches (see Appendix 1 for a list of terms in these topics). The sample models in Appendix A show that the topic model algorithms were, indeed, picking up thematic similarities between plays and shared preoccupations of characters.

This paper focuses on Shakespeare's second tetralogy, the Henriad. The Henriad is a group of four plays (*Richard II*, *1 & 2 Henry IV*, and *Henry V*) primarily focused on one person: King Henry V, who appears in the final three plays. Scholars have long discussed the trajectory of Henry's character (see, for instance, Dickinson 1961, Schell 1970, Frazer 2013): as Henry himself explains, he will pretend to be debauched and irresponsible—his ascension to the throne and reformation will be even more powerful than if he had been good all along: "My reformation, glittering over my fault,/Shall show more goodly and attract more eyes/Than that which hath no foil to set it off" (*1 Henry IV* 1.2.212-215).[5] At the end of *2 Henry IV*, Prince Hal becomes King Henry, and announces his new self: "Presume not that I am the thing that I was / For God doth know, so shall the world perceive, / That I have turn'd away my former self" (5.5.56-58). Even as a carousing young prince, Hal has planned his transformation into responsible king, which he puts in motion at his coronation.

Falstaff, a drinking buddy and father figure for young Prince Hal, also appears in multiple plays: in *1* and *2 Henry IV*, he traces a trajectory opposite of Hal's: he begins as a boisterous drunk, a dishonourable yet happy knight content with his beer and

[4] We use single quotation marks to mark terms taken from our topic model and double quotation marks to indicate quotations from the plays and critical sources.

[5] All line numbers from Shakespeare's plays are taken from The Riverside Shakespeare, 2nd ed. (Evans 1997).

his women, and ends by being rejected by Hal upon his accession to the throne. Falstaff is also mentioned in *Henry V*, as the other characters reminisce fondly about his drinking and womanizing after his death. Falstaff was so popular that after Shakespeare killed him at the end of *2 Henry IV*, he was brought back for a comedy, *The Merry Wives of Windsor*, set in a different reality than Shakespeare's histories. As the old story goes, Queen Elizabeth enjoyed Falstaff's character so much, it was she who requested he return to the stage (Crane 1997, 3).

Our topic modeling demonstrates that Prince Hal (Henry V) is a qualitatively different character in the plays in which he appears, whereas Falstaff remains the same. This might be counter-intuitive, as Prince Hal/Henry V is indeed the same historical figure that Shakespeare dramatizes in multiple plays, yet Falstaff is a fictitious character Shakespeare created to suit two different genres. Indeed, while Shakespearean scholars might be willing to accept a reformed Henry V who is different from his younger self, stating that Falstaff of *Merry Wives* is the same Falstaff from the history plays is a more contentious claim.

From Prince Hal to Henry V

Henry V appeared prominently in three topics that our model created. For instance, in a topic that included words such as king, lord, duke, and God as the most probable words, which we'll call the "kingly topic," Henry in the play *Henry V* had a 34% chance of belonging (see Appendix A for a list of words in the "kingly topic"). A few other characters from this play had much higher probabilities of being in this topic—for instance, the Duke of Bedford or the Duke of Gloucester. This, however, seems to be expected: these characters speak far fewer lines than the hero of the play: in this case, six and five lines respectively. If you speak only fifty words or so, then you are more likely to be on point or speaking about the same subject. For characters who speak more lines, such as Henry, they are more likely to discuss varying ideas. In *Henry V*, the title character speaks over a thousand lines: an amount of speech exceeded only by Iago and Hamlet in all of Shakespeare's plays.[6] In those over-a-thousand lines,

[6] This number of lines per character is taken from Shakespeare's Words (Crystal and Crystal 2008), which is based on the New Penguin Shakespeare series.

Henry is likely to be talking about kings, gods, and lords in *Henry V*. However, in the prequels, the *Henry IV* plays, his younger self, Prince Hal, is nowhere near as likely to talk about such noble and kingly terms. For this topic, Prince Hal when he first appears (in *1 Henry IV*) is only 5% as likely to discuss these terms, whereas by *2 Henry IV*, his likelihood of discussing these ideas increases marginally to 8%, which might be accounted for by his reformation at the end of the play. Upon hearing that he has inherited the throne, Henry begins to use more formal diction immediately: "The tide of blood in me / Hath proudly flowed in vanity till now; / Now it doth turn and ebb back to the sea / Where it shall mingle with the state of floods, / And flow henceforth in formal majesty" (*2 Henry IV* 5.2.129-33). Henry at once calls the parliament to session so he can be crowned and begin the business of governing.

When Henry becomes king, he starts talking more like his father, King Henry IV. When King Henry IV is on the throne, he uses terms in the "kingly topic" at almost the same rate as his son does when he is crowned. In *1 Henry IV*, the titular king has a 34% probability of belonging to this topic, just like his son, King Henry V, in his title play. In *2 Henry IV*, however, Henry IV's probability of belonging to the "kingly" topic drops, although not precipitously: he is only 21% likely to belong. In *2 Henry IV*, he's more concerned with the sin of how he took the crown from the previous king, Richard II. As King Henry IV says to his son, "God knows, my son, / What by-paths and indirect crooked ways / I met this crown, and I myself know well / How troublesome it sat upon my head" (*2 Henry IV* 4.5.183-86). The change in this topic model could reflect how before his death, King Henry IV begins to turn away from the business of actually ruling, and instead thinks of past regrets.

We are not arguing that having a high probability of being associated with the "kingly topic" makes a character more likely to be a ruler. Often, however, kings are around the 30–40% mark of probability. For instance, in *Richard II*, King Richard II's speeches have a 38% chance of belonging to this topic. In plays beyond the Henriad, Richard III and Henry VI fall into this range. Aspirants to the throne do at times, too: Jack Cade, leader of Cade rebellion, exhibits similar traits. Nobles can also speak in this register: Queen Margaret in the *Henry VI* plays is an example of this. Having a higher percentage in this category, however, does not indicate that a character is

necessarily more kingly—as we discussed, characters with only a few lines can rank disproportionately high in any given topic. That said, these data clearly show that Prince Hal is not interested in the language of rulership before he comes to the throne, which is exactly what he tells the audience.

Prince Hal tells his audience that we should consider his youth a "foil" to his mature age (*1 Henry IV* 1.2.215). Shakespeare's dramaturgy puts Hal in natural comparison to a number of other characters: to his father, the previous king; to his rival, Hotspur (who was also named Henry); to his brother, Prince John; we can even compare him to his son, Henry VI, the star of Shakespeare's first tetralogy. Shakespeare wrote the Henry VI group of plays first—they were big hits that prompted him to write more history plays. Then, like George Lucas did with *Star Wars*, he went back to write the prequels, the Henriad (also called the second tetralogy). At the end *of Henry V*, Shakespeare invites the audience to think of Henry V in relation to his son—a story early modern audiences would have already known from Shakespeare's earlier plays. As the Epilogue relates, King Henry V was "This star of England. Fortune made his sword; / By which the world's best garden he achieved, / And of it left his son imperial lord. / Henry the Sixth, in infant bands crowned king / Of France and England, did this king succeed; / Whose state so many had the managing, / That they lost France, and made his England bleed" (*Henry V* Epilogue.6-12). Henry V won England great glory (and part of France); his son lost it all.

Topic modeling prompts us to find foils for Henry beyond those set up by the structure of Shakespeare's plays. In our model of *Henry V*, the character who speaks most like Henry on the "kingly topic" is not actually a historical character at all: it is the chorus. In *Henry V*, the chorus serves to introduce the play, to tell the audience about offstage events, and to interpret the events presented in a way that manipulates the audience response. What this topic model suggests is that the chorus is equally invested in kingship and how it functions as Henry V is when he becomes king.

Up to this point, we've been focusing on the words most likely to be in a given topic: for instance, in the "kingly topic," the terms 'king' and 'lord' have two percent probability of appearing in this topic. Although this might seem low, in our model,

that suggests that those are among the most popular words in this topic. Both 'king' and 'lord' can be used as a form of address, which means that they might be among the most frequent words in the history plays. However, this model is based not on the frequency of terms but on the probability of a given word appearing. And while some of the other words in this topic are indeed terms of address, such as queen and majesty, this topic is not entirely based on titles (for instance, 'friar,' which appears prominently in other topics, does not appear in this one). Some of the other words in this topic spell out the importance of kingship: 'war,' 'sword,' and 'soldier' suggest the martial aspects of this topic, whereas 'good,' 'noble,' and 'peace' suggest the qualities that a good king should hold. The appearance of terms such as 'crown,' 'sovereign,' and 'royal' reinforce our calling this the "kingly topic." 'England' and 'English' appear in this topic—unsurprising for plays about the Kings of England—but 'France' appears too, which highlights the importance of England's greatest rival and the English holdings on French soil.

Naturally, a number of the kings' names appear in the "kingly topic," including Henry, Gloucester, and Richard. What we might not expect is for 'Edward' to appear on this list. The name "Edward" appears in only two topics. The "kingly topic" (discussed here) and another topic derived primarily, if not entirely, from stage directions (see Appendix A). Although King Edward IV and his grandson prince Edward both appear briefly in *Richard III* and could be responsible for the appearance of this word in this topic, it is equally likely that the king of a century earlier, Edward III, is responsible for evoking this name. Edward III was the last king before the War of Roses: before the era in which eight of Shakespeare's history plays take place.[7] All of the kings in the Henriad proudly trace their ancestry back to Edward: they can claim kingship only because they share "Edward's sacred blood" (*Richard II* 1.2.17). Henry V's legal claim to the throne is based on this heritage: when the ambassador to France tries to flatter him, he calls Edward Henry's "great predecessor, King Edward the Third" (*Henry V* 1.2.248). What this topic suggests is that kingship in Shakespeare's War of the Roses

[7] There is an early modern history play called *Edward III* that some scholars claim was written in part but Shakespeare, but like other apocryphal plays, it was not included in our corpus. See Sams 1996, Bruster 2015, and Freebury-Jones 2016.

plays is as much about tracing inheritance to a now-legendary past king as it is being a noble warrior. Using a topic model based on probability instead of frequency allows us to find areas of comparison that might otherwise be overlooked, such as the importance of Edward III (a king who never appears onstage) to Shakespeare's history plays. Of course, the importance of Edward III to the Henriad has already been explored in scholarship, to the point where the *Riverside Shakespeare* gives the title "Edward's Sacred Blood" to its family tree (Evans 1997, 630–631).

We are not, of course, the first to discuss kingship in *Henry V*, nor are we the first to suggest that Prince Hal's character changes. Our methods, however, have led us to make new connections and draw new conclusions: these topic models quantify Hal's reformation, lead us to draw comparisons between Henry V and the Chorus of that play, and induce us to consider the importance of characters such as Edward III who do not even appear in the Henriad.

Is Falstaff Falstaff?

In *1 & 2 Henry IV*, Falstaff is a boisterous, debauched, drunk. Hal must reject Falstaff in order to achieve his reformation and become king: the climactic moment of *2 Henry IV* is Henry V's cold proclamation to Falstaff, "I know thee not, old man" (5.5.47). With his speech during his coronation procession, Henry V rejects his ill-spent youth and repudiates Falstaff as the personification of his earlier days.

Although he dies offstage between *2 Henry IV* and *Henry V*, Falstaff's reputation extends beyond his death. Fluellen, a soldier, recounts Falstaff's rejection by Prince Hal: "Harry Monmouth [Prince Hal], being in his right wits and his good judgments, turned away the fat knight with the great belly-doublet: he was full of jests, and gipes, and knaveries, and mocks; I have forgot his name" (*Henry V* 4.7.46-50). His interlocutor, Gower, supplies the missing name and names the ghost that metaphorically haunts *Henry V*: "Sir John Falstaff" (4.7.51). Falstaff is a larger-than-life character whose presence looms beyond those plays in which he actually appears.

Although Falstaff is a fictional character, not taken from Shakespeare's history book sources, he is based on historical figures and literary stock characters. Falstaff

combines the classical stock figure of the Miles Gloriosus (boastful yet cowardly knight) with the stock character of the glutton and the fool. Shakespeare takes great pains to disavow that Falstaff is connected to Sir John Oldcastle, a historical friend of Prince Hal's: as the epilogue to *2 Henry IV* states, "Falstaff shall die of a sweat, unless already a' be killed with your hard opinions; for Oldcastle died a martyr, and this is not the man" (Epilogue.30-32). Shakespeare's claim that Falstaff is not Oldcastle, of course, only serves to draw the comparison for the audience (see, for instance, Taylor 1985 and Melchiori 1994).

As an amalgam of historical and literary sources who Shakespeare invented to suit his dramatic needs, Falstaff's character might be expected to be adaptable depending on the contexts in which he appears. Our topic model, however, suggests that Falstaff is Falstaff, whether he be in the history plays in which he originally appeared, or in the comedy spin-off, *Merry Wives of Windsor*, in which he starred. *Merry Wives* occurs in a timeline that seems irreconcilable to the Henriad: scholars debate if *Merry Wives* is a prequel or sequel (see, for instance, Melchiori 1994). Most recently, Kay Stanton (2015) has argued that *Merry Wives* takes place in a feminist parallel universe. Our topic model suggests that there is something ultimately Falstaffian that an algorithm might capture: Falstaff is, indeed, Falstaff: across plays, across places, and across genres.

Falstaff was primarily written by the same author, Shakespeare. Of course, despite Shakespeare's predominant hand in these plays, every early modern play is, to a degree, an act of collaborative authorship. Will Kempe, the actor who played Falstaff, would have contributed to each performance and improvised certain elements, which may or not be captured in the printed version. Plays could be revised by other playwrights working for the companies or altered, to varying degrees, by printers or compositors. Falstaff, then, can be considered a collaborative creation, and we cannot know how much the printed version reflects this. Topic modeling, however, allows us to see how much the published (and later, encoded) version of Falstaff in each play is the same or different.

Of the 49 topics that our model generated, only two of our topics featured Falstaff from all three plays: the first was a list that favored general terms of address ("Sir,"

"master," 'come,' 'good,' 'well,' 'man,' 'mistress'…); the second was a grab-bag of words ('love,' 'thou,' 'shall,' 'man,' 'come,' 'would') that didn't reflect a particularly theme or play. Turning to "topics" that seem, at first, not to be topic-based can be fruitful: the field of stylometrics has shown the value of looking at often-overlooked words to determine authorship (Juola 2013).[8] The general nature of these words doesn't mean the results for this topic should be discarded: indeed, these non-thematic topics suggest that there the statistical model sees a similarity between Falstaff that goes beyond his love of women and sack—that is to say, although a humanist close reader might not associate these words with each other or Falstaff, this algorithm does.

The fact that Falstaff from all three plays appears in only two of our forty-nine topics could seem like a relatively low occurrence: however, Falstaff (in all plays) occurs only in four topics created by our model. **Table 1** demonstrates that Falstaff from *1 Henry IV* is 75% likely to share a topic with his later self in *2 Henry IV* (unlike Prince Hal, who does not show the same similarity with his later self). But perhaps most surprisingly, Falstaff from *The Merry Wives of Windsor* is just as likely to share a topic from the same character from the history play. There is no topic where Falstaff from one play appears alone without Falstaff from another. Even though *The Merry Wives of Windsor* takes place in a different universe from the history plays, he is basically the same Falstaff. The Falstaff of the history plays is a fictionalized character living in a true historical setting, whereas Falstaff of *Merry Wives* is part of a realistic and not historicized English town. Despite the genre-based differences between these plays, Falstaff is Falstaff.

Table 1: Topics where Falstaff appears (see also Appendix A).

Topic	*1 Henry IV*	*2 Henry IV*	*Merry Wives of Windsor*
4	*	*	
16	*	*	*
32	*	*	*
41		*	*

[8] Indeed, topic modeling has been expanded to consider authorship (stylometrics/stylometry) in some cases (see the oft-cited article by Rosen-Zvi, Griffiths, Steyvers, and Smyth, 2004).

Conclusions

Topic modeling, when paired with thoughtfully encoded texts, can lead us to ask questions that we might not otherwise have asked. However, without an understanding of these texts, topic modeling can reveal precious little. Even productive topic models, such as this one that led us to reconsider the nature of Prince Hal and Falstaff, will not always offer fruitful comparisons. For instance, one of the topics that this model generated had "shall, would, make, well, sir, us, and yet" as some of the most probable words. While we did not find this an interesting avenue for further research, though, perhaps, a linguist working on modals would. The topic model that we created did not provide the questions to us: it gave us comparisons, themes, and hints that led us to our research questions. To put it another way: this topic model provides a distant reading (following Jockers 2013 and Moretti 2013) of Shakespeare's plays, which can be supported by more traditional literary analysis and textual evidence and can lead to new research questions.

Ultimately, our research demonstrates the fruitfulness not just of topic modeling, but of topic modeling a dramatic corpus by character. The first major conclusion of this research, that Prince Hal is not the same as Henry V, is one that the character himself tells us: yet our model shows that Hal is categorically and quantifiably different (not just claiming to be so). Our second major conclusion—that Falstaff's character does not change over multiple plays—adds a new form of evidence to be considered in a centuries-long debate in Shakespeare studies. Indeed, it is our hope that this algorithmically-informed analysis of Shakespeare's plays is not taken as a final answer, but, instead, as a prompt. As this research demonstrates, topic modeling plays with attention to each speaker opens the door for new comparisons that expand on and reassess previous critical interpretations.

Additional File

The additional file for this article can be found as follows:

- **Appendix 1.** Sample Results. DOI: https://doi.org/10.16995/dscn.295.s1

Acknowledgements

We would like to thank our reviewers for their thoughtful feedback. Thanks also to Constance Crompton for her encouragement to pursue this topic.

Competing Interests

The authors have no competing interests to declare.

References

Abrams, M. H. 1999. *A Glossary of Literary Terms*. Boston: Heinle & Heinle—Thompson.

Bakhtin, M. M. 1981. *The Dialogic Imagination: Four Essays*. Edited by Michael Holquist. Translated by Caryl Emerson and Michael Holquist. Austin and London: University of Texas Press.

Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55(4): 77–84. DOI: https://doi.org/10.1145/2133806.2133826

Blei, David M., Andrew Y. Ng, and **Michael I. Jordan.** 2003. "Latent Dirichlet Allocation." *The Journal of Machine Learning Research* 3: 993–1022.

Bloom, Harold. 1998. *Shakespeare: The Invention of the Human*. New York: Riverhead—Penguin Putnam.

Bruster, Douglas. 2015. "Shakespeare's Pauses, Authorship, and Early Chronology." *Studia Metrica et Poetica* 2(5): 25–47. Accessed September 16, 2017. https://ojs.utlib.ee/index.php/smp/.

Carlson, Marvin. 1992. "Theater and Dialogism." In *Critical Theory and Performance*, edited by Janelle G. Reinelt and Joseph R. Roach, 313–23. Ann Arbor: The University of Michigan Press.

Coleridge, Hartley. 1851. *Essays and Marginalia*. London: Edward Moxon. Accessed September 16, 2017. https://archive.org/details/essaysandmargin01colegoog.

Crane, David, ed. 1997. *The Merry Wives of Windsor*. Cambridge: Cambridge University Press.

Crystal, David, and **Ben Crystal.** 2008 *Shakespeare's Words*. Last modified. Accessed September 16, 2017. http://shakespeareswords.com.

Dickinson, Hugh. 1961. "The Reformation of Prince Hal." *Shakespeare Quarterly* 12: 33–46. DOI: https://doi.org/10.2307/2867269

Duhaime, Douglas. 2014. "Classifying Shakespearean Drama with Sparse Feature Sets." Accessed December 5, 2017. http://douglasduhaime.com/posts/classifying-shakespearean-drama-with-sparse-feature-sets.html.

Estill, Laura, and **Luis Meneses.** Digital Acting Parts. Accessed September 16, 2017. http://digitalactingparts.tamu.edu/dap/.

Evans, G. Blakemore, ed. 1997. *The Riverside Shakespeare.* 2nd ed. Boston: Houghton Mifflin.

Flanders, Julia. 2012. "Modeling Scholarship." Presentation at Knowledge Organization and Data Modeling in the Humanities Workshop, Brown University. March 15. https://datasymposium.wordpress.com.

Frazer, Paul. 2013. "Itinerant Identities: England Walking Low in *Henry IV*." *Shakespeare* 9(1): 1–20. DOI: https://doi.org/10.1080/17450918.2012.705881

Freebury-Jones, Darren. 2016. "Kyd and Shakespeare: Authorship, Influence, and Collaboration." PhD diss., Cardiff. Accessed September 16, 2017. http://orca.cf.ac.uk/91745/.

Goldstone, Andrew, and **Ted Underwood.** 2014. "The Quiet Transformation of Literary Studies: What Thirteen Thousand Scholars Could Tell Us." *New Literary History* 45(3): 359–84. DOI: https://doi.org/10.1353/nlh.2014.0025

Henrichs, Amanda. 2017. "Topic Modelling the *Sonnets*." Seminar contribution (unpublished), Shakespeare Association of America, "Shakespeare by the Numbers," Atlanta, 6 April.

Hylton, Jeremy, ed. "The Complete Works of William Shakespeare." *The Tech. MIT*. Accessed September 16, 2017. http://shakespeare.mit.edu/.

Jockers, Matthew L. 2013. *Macroanalysis: Digital Methods & Literary History*. Champaign, IL: University of Illinois Press.

Johnson, Eric, ed. *Open Source Shakespeare*. George Mason University. Accessed September 16, 2017. www.opensourceshakespeare.org.

Juola, Patrick. 2013. "How a Computer Program Helped Reveal J. K. Rowling as Author of *A Cuckoo's Calling*." *Scientific American*. 20 August. Accessed September 16, 2017. http://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/.

Melchiori, Giorgio. 1994. *Shakespeare's Garter Plays: Edward III to Merry Wives of Windsor*. Newark: University of Delaware Press.

Moretti, Franco. 2013. *Distant Reading*. London and New York: Verso.

Nalisnick, Eric T., and **Henry S. Baird.** 2013a. "Character-to-Character Sentiment Analysis in Shakespeare's Plays." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 479–83. Sofia, Bulgaria: Association for Computational Linguistics. http://aclanthology.info/pdf/P/P13/P13-2085.pdf.

Nalisnick, Eric T., and **Henry S. Baird.** 2013b. "Extracting Sentiment Networks from Shakespeare's Plays." In *ICDAR 2013 12th International Conference on Document Analysis and Recognition*, 759–61. http://www.icdar2013.org/docs/detailedprogram.html.

Niles, Rebecca, and **Michael Poston.** 2016. "Re-modeling the Edition: Creating the Corpus of Folger Digital Texts." In *Early Modern Studies after the Digital Turn*, edited by Laura Estill, Diane Jakacki, and Michael Ullyot, 119–46. Tempe and Toronto: Iter and Arizona Center for Medieval and Renaissance Studies.

Palfrey, Simon, and **Tiffany Stern.** 2007. *Shakepeare in Parts*. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199272051.001.0001

Rackin, Phyllis, and **Evelyn Gajowski.** 2015. "Introduction: A Historical Survey." In *The Merry Wives of Windsor: New Critical Essays*, edited by Gajowski and Rackin, 1–24. London: Routledge.

Řehůřek, R., and **P. Sojka.** 2010. "Software Framework for Topic Modelling with Large Corpora." In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Malta: Valletta.

Rhody, Lisa M. 2012. "Topic Modeling and Figurative Language." *Journal of Digital Humanities* 2(1). Accessed September 16, 2017.

http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/.

Rosen-Zvi, Michal, Thomas Griffiths, Mark Steyvers, and **Padhraic Smyth.** 2004. "The Author-Topic Model for Authors and Topics." In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, edited by Christopher Meek and Joseph Halpern. Arlington, VA: AUAI Press. Accessed September 16, 2017. http://psiexp.ss.uci.edu/research/papers/uai04_v8.pdf.

Sams, Eric, ed. 1996. *Shakespeare's Edward III*. New Haven and London: Yale University Press.

Schaefer, Kayla Hope. 2015. "Document Clustering with Nonparametric Hierarchical Topic Modeling." MA thesis, University of Texas at Austin.

Schell, Edgar T. 1970. "Prince Hal's Second 'Reformation.'" *Shakespeare Quarterly* 21: 11–16.

Schmidt, Benjamin J. 2012. "Words Alone: Dismantling Topic Models in the Humanities." *Journal of Digital Humanities* 2(1). Accessed September 16, 2017. http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/.

Schmidt, Benjamin J. 2016. "Do Digital Humanists Need to Understand Algorithms?" In *Debates in the Digital Humanities 2016*, edited by Matthew K. Gold and Lauren F. Klein. Minnesota: University of Minnesota Press. Accessed September 16, 2017. http://dhdebates.gc.cuny.edu/debates/text/99.

Spevack, Martin. 1974. *The Harvard Concordance to Shakespeare*. Cambridge, MA: Belknap Press of Harvard University Press.

Stanton, Kay. 2015. "Shakespeare's Quantum Physics: *Merry Wives* as Feminist 'Parallel Universe' of *Henry IV, Part 2*." In *The Merry Wives of Windsor: New Critical Essays*, edited by Evelyn Gajowski and Phyllis Rackin, 84–95. London: Routledge.

Taylor, Gary. 1985. "The Fortunes of Oldcastle." *Shakespeare Survey* 38: 85–100.

Van Rossum, Guido. 1995. "Python Tutorial." In *Report CS-R9526*. Amsterdam: Centrum voor Wiskunde en Informatica (CWI).

How to cite this article: Estill, Laura and Luis Meneses. 2018. "Is Falstaff Falstaff? Is Prince Hal Henry V?: Topic Modeling Shakespeare's Plays." *Digital Studies/Le champ numérique* 8(1): 1, pp. 1–22, DOI: https://doi.org/10.16995/dscn.295

Submitted: 19 October 2016 **Accepted:** 09 May 2017 **Published:** 23 January 2018

Digital Studies/Le champ numérique is a peer-reviewed open access journal published by Open Library of Humanities. **OPEN ACCESS**