Statistical Stylometrics and the Marlowe-Shakespeare Authorship Debate

Neal Fox[1]
Department of Cognitive, Linguistic & Psychological Sciences
Brown University

Omran Ehmoda and Eugene Charniak
Department of Computer Science
Brown University

{npfox, oehmoda, ec} @cs.brown.edu

I. Authorship Attribution Paradigm

Historians, literary scholars, psychologists, and – more recently – computational linguists have long sought a reliable methodology for analyzing texts to determine the identity of their author. Since at least the late nineteenth century (see Mendenhall, 1887; Mascol, 1888a/b), one tool used in the investigation of authorship has been the extraction of statistical tendencies from the documents and comparison of these data in order to group the documents appropriately. Underlying the search for such a methodology is the critical assumption that some statistically quantifiable characteristic or set of characteristics inherent in a single author's use of written language could be isolated that would be consistent across works by that author, but differ between different authors. Thus, the feature(s) could be used as a sort of fingerprint to distinguish between works by distinct authors and to identify the likely author of an anonymously published work. This endeavor is generally referred to as stylometry.

With the advent of natural language processing and machine learning techniques in recent decades, the field has advanced greatly, and numerous researchers have addressed many specific

[1] NF thanks John Bonvillian (Department of Psychology, University of Virginia) for introducing him to the problem, for encouraging him to pursue the question, and for frequent conversations regarding the topic and the research. NF is funded by an NSF Graduate Research Fellowship. The authors also thank Micha Elsner for many helpful suggestions and conversations that contributed to the development of this project.

questions and disputes about authorship across different genres, from the Federalist Papers to the New Testament to email and blog postings. These researchers have suggested many different feature sets as potential fingerprint markers, from word and sentence length distributions to punctuation mark and function word frequency. Indeed, Koppel et al (2009) and Grieve (2007) completed comprehensive studies to determine which feature sets are most accurate in predicting the authorship of works whose author is masked (see also Stamatatos, 2009 for a review of current models). Furthermore, researchers have used different techniques for learning which features are strong predictors and which are just noise, from simple distance metrics to complex machine learning algorithms. We consider all variants of this question to fall into a class of computational models and tasks that we will simply call the authorship attribution paradigm.

It should be noted that many other degrees of classification besides the specific author identity are of interest to forensic style researchers, and a great number have been considered, such as gender (e.g. Koppel et al, 2002; Argamon et al, 2003), age (Burger & Henderson, 2006; Schler et al, 2006), and native language of the author (Koppel et al, 2005).

II. Shakespearean Authorship Dispute

William Shakespeare is touted by scholars across the world as the greatest dramatist, even writer, not only in the English language, but in any language (see e.g. Wells, 1997). Shakespeare is a sort of household celebrity; literate English-speakers of all ages and backgrounds have read or can identify his works. However, many scholars dispute that he was the man most think him to have been. There is a body of evidence that has raised questions about whether William Shakespeare actually wrote or even could have written the works attributed to him. Individuals who contest the traditional view of Shakespearean authorship are sometimes referred to as *anti-Stratfordians* for their denial that the works attributed to Shakespeare were

written by the man of that name who we know from historical records lived in Stratford-upon-Avon, England in the late sixteenth century and early seventeenth century.

Various pieces of evidence for this denial have been cited by anti-Stratfordians, including that Shakespeare may have been uneducated, untraveled, and illiterate or barely literate (see, e.g. Bethell, 1991; Nelson, 2004). He did not leave behind a single letter to anyone, and he lists no books or papers in his will nor does he mention his vast still-unpublished collection of plays (Price, 2001). In this era, it was not uncommon to have long delays between the first known performances and the actual publications of plays, which increases the uncertainty surrounding the authorship of the disputed Shakespearean canon and many contemporary works controversially or tentatively attributed to other playwrights (Harbage, 1989). These, among hundreds of other historical and literary facts, are offered as evidence suggesting that the Shakespearean authorship question ought to be at least examined. Still, these pieces of evidence are considered conspiracy theories and/or classist notions of intellectual capability and subjected to a vicious resistance from *Stratfordians* intent on defending the "Bard of Avon's" honor (see, e.g. McMichael & Glenn, 1962).

Specifying the case for either side of the authorship debate is beyond the scope of this paper, and we leave it to the reader to investigate any historical or literary data. We instead intend to focus on the exploitation of statistical properties of the language in the texts from the period to investigate the question of authorship. Importantly, this is as much an exercise in development of an authorship attribution model as it is an attempt to find a suitable resolution to an old debate. We begin our investigation under the aforementioned assumption that there are properties of an author's style which are consistent between his or her works, and that those properties will often vary between different authors.

There has been substantial previous work on the question of Shakespearean authorship using the principles of stylometry. Mendenhall (1901) examined the word length frequency distribution of different authors, showing that, in fact, there was consistency along this distribution across works by an author, and it is also true that sometimes authors had very different word length frequency distributions. For example, Shakespeare used four-letter words more frequently than three-letter words, whereas Francis Bacon (a candidate who has been discussed as a possible author of the Shakespearean canon) used three-letter words much more than four-letter words. In the same study, Mendenhall claimed that the word-length frequency distribution of Christopher Marlowe (a popular candidate in the authorship debate) "agrees with Shakespeare as well as Shakespeare agrees with himself." Williams (1975), however, demonstrated that such comparisons were inherently across genres (verse vs. prose), and that the study was therefore not properly controlled; he illustrates a false positive and false negative result using verse and prose by Elizabethan contemporary Philip Sidney. Slater (1988) studied communities of vocabulary (i.e. groups of words that occur together in works) and rare words in the anonymously published 1596 play *Edward III* and compared these tendencies with works in the Shakespearean canon and works of other authors suspected of having written that much disputed play, finally concluding that it was very likely written by the same individual who wrote the works commonly attributed to Shakespeare. Merriam (1996, 1998), Matthews and Merriam (1993) and Merriam and Matthews (1994) used machine learning and statistical techniques such as neural networks, principle component analysis, and multilayer perceptrons to learn linguistic properties of the works thought to be by Shakespeare, Marlowe, John Fletcher (a contemporary playwright and suspected collaborator on some of the later works attributed to Shakespeare), and many other authors of the time; their results served to cast doubt on the authorship of several

specific works that appeared to be anomalous to the Shakespearean canon. Craig and Kinney (2009) also did a large multivariate analysis of authorship with respect to the Shakespearean question. Like Merriam and colleagues, this analysis, while extensive, scrutinized only a few of the works attributed to Shakespeare.

Many other studies have attempted such stylometric analyses. This paper however, is an attempt to examine the hypothesis that Christopher Marlowe wrote the *entirety* of the Shakespearean canon.

The "Marlovian Theory," briefly, is that the influential dramatist Christopher Marlowe did not, as historical records state, die on May 30, 1593, but he actually escaped death and continued writing in exile with William Shakespeare's name as a front (see, e.g. Schoenbaum, 1991). Many modern scholars have noted the substantial similarities between Marlowe's and Shakespeare's works – in language and in content – and it is inescapable that the blank-verse style pioneered by Marlowe (Shaw, 2007) was a major influence on Shakespeare's style (Ackroyd, 2005). The Marlovian Theory takes these similarities a step further, asserting that the two men were actually one, and Shakespeare simply "Late Marlowe."

However, the theory that Marlowe was the author of all, or even most, of the works commonly attributed to Shakespeare, while attracting a sizeable number of followers, has, until now, not to our knowledge received a close analysis using computational tools and a variety of potentially relevant stylistic features (Pinksen, 2008, for example, extended the results of Mendenhall, 1901, but still looked at a relatively limited feature set). Furthermore, any such authorship attribution model as would emerge from this endeavor ought to be validated by testing a large body of works by a number of contemporary Elizabethan authors who are relatively unlikely to be touted as the true author of the Shakespearean canon.

Under the assumption that an author has a somewhat consistent distribution of some statistically measurable, stylistically relevant feature set, we will develop two very different models of authorship attribution and test them to determine whether the style of Marlowe and Shakespeare are statistically more similar than between, for instance, Marlowe and Ben Jonson (a contemporary playwright) or Shakespeare and Jonson. Stylistic similarity will be measured based on likelihood of classification confusions between the canonical divisions between authors. In other words, a much higher probability of confusions between canonical Shakespeare and canonical Marlowe (for instance, *Henry VI, Part I* being consistently classified as a work of Marlowe or *Edward II* being classified as more likely to be a work of Shakespeare) than confusions between canonical Shakespeare and canonical Jonson would indicate that the author(s) of the works in the Marlowe canon and the Shakespeare canon are difficult to distinguish, and possibly the same individual.

It is very important to note that any method like ours that "trains on" data is learning the author-labeled statistical distributions of a set of works that has been pre-classified. It is therefore highly dependent on the assumptions of the individuals doing the pre-classification. For instance, as the experimenters, we are suggesting to the model that all the works it trains on in a given category are by the same person. We are not allowing the model to "find" Fletcher, for instance, if one of the works attributed to Shakespeare is actually by Fletcher. We attempt to correct for this with a different kind of methodology introduced later in this article called "unsupervised clustering." Using this methodology, there is no training. We simply tell the model which features to measure and set it free to find the groupings (with guidance from the experimenter as to the number of groupings) that minimize the total variance in the corpus.

Furthermore, it is important to note that our model can only at best imply that two different authors are *more similar* than any other pairing of authors. It is impossible to distinguish whether two very similar categories in the corpus are sets of works by the *same* author, or simply works by two very *similar* authors. This is certainly a major hurdle of any authorship attribution method that compares statistical distributions between works and authors in the way we do.

III. Corpus

We chose three other prominent contemporary dramatists with a substantial canon besides Shakespeare and Marlowe. We strived to obtain at least 200,000 words for each of the authors by seeking out machine-readable texts freely available on the internet that could be easily stripped of any language that was not part of a line spoken by an actor. We removed, as best we know, character labels, prologues (unless they were spoken by a character), stage directions, act and scene breaks, footnotes, etc. Though some research has shown that punctuation can be useful in authorship attribution (e.g. Chaski, 2001; Mascol, 1888a/b), any punctuation marks were skipped over in the analysis because of the chance that they were added by the editors rather than the authors, even centuries later, as we did not have access to the original manuscripts of every work. We predict that punctuation would be more likely than words to be altered by an editor, modern or not. Table 1 lists by author and year the titles of the texts included in our corpus. Table 2 lists the authors included and what proportion of the corpus their work makes up.

	Author	Title	Year
1	Chapman	The Blind Beggar of Alexandria	1596
2	Chapman	An Humorous Day's Mirth	1597
4	Chapman	All Fools	1601
5	Chapman	May Day	1602
3	Chapman	Sir Giles Goosecap	1602
6	Chapman	The Gentleman Usher	1602
7	Chapman	Bussy D'Ambois	1604

8	Chapman	The Widow's Tears	1604
9	Chapman	Monsieur D'Olive	1605
10	Chapman	Conspiracy and Tragedy of Charles Duke of Byron	1608
11	Chapman	The Revenge of Bussy D'Ambois	1610
1	Jonson	A Tale of a Tub	1596
2	Jonson	The Case is Altered	1597
3	Jonson	Every Man in His Humour	1598
4	Jonson	Cynthia's Revels, or The Fountain of Self-Love	1600
5	Jonson	Poetaster	1601
6	Jonson	Volpone	1606
7	Jonson	The Alchemist	1610
8	Jonson	Catiline His Conspiracy	1611
9	Jonson	Love Restored	1612
10	Jonson	Bartholomew Fair	1614
11	Jonson	The Devil is an Ass	1616
12	Jonson	The Staple of News	1626
13	Jonson	The New Inn, or The Light Heart	1629
14	Jonson	The Magnetic Lady	1632
1	Marlowe	Dido	1586
2	Marlowe	I Tamburlaine the Great	1587
3	Marlowe	II Tamburlaine the Great	1588
4	Marlowe	The Jew of Malta	1589
5	Marlowe	Doctor Faustus	1592
6	Marlowe	Edward II	1592
7	Marlowe	The Massacre at Paris	1593
1	Middleton	The Family of Love	1603
2	Middleton	The Phoenix	1604
3	Middleton	A Trick to Catch the Old One	1605
4	Middleton	The Five Gallants	1607
5	Middleton	No Wit, No Help Like a Woman's	1611
6	Middleton	The Second Maiden's Tragedy	1611
8	Middleton	A Chaste Maid in Cheapside	1613
7	Middleton	The Witch	1613
9	Middleton	Hengist, King of Kent	1618
1	Shakespeare, Early	I Henry VI	1590
2	Shakespeare, Early	II Henry VI	1590
4	Shakespeare, Early	III Henry VI	1591
3	Shakespeare, Early	The Life and Death of King John	1591
6	Shakespeare, Early	The Comedy of Errors	1592
7	Shakespeare, Early	The Taming of the Shrew	1592
5	Shakespeare, Early	The Tragedy of King Richard III	1592
8	Shakespeare, Early	The Two Gentlemen of Verona	1593
9	Shakespeare, Early	Titus Andronicus	1594
10	Shakespeare, Early	Love's Labour's Lost	1595
11	Shakespeare, Early	Richard II	1595
13	Shakespeare, Early	A Midsummer-Night's Dream	1596
14	Shakespeare, Early	Romeo and Juliet	1596
12	Shakespeare, Early	The Merchant of Venice	1596
15	Shakespeare, Early	I Henry IV	1597
16	Shakespeare, Early	II Henry IV	1597

17	Shakespeare, Early	The Merry Wives of Windsor	1597
18	Shakespeare, Early	Much Ado about Nothing	1598
19	Shakespeare, Early	As You Like It	1599
20	Shakespeare, Early	Henery V	1599
21	Shakespeare, Early	Julius Caesar	1599
23	Shakespeare, Early	Hamlet	1601
22	Shakespeare, Early	Twelfth Night, or What You Will	1601
24	Shakespeare, Late	Troilus and Cressida	1602
25	Shakespeare, Late	All's Well That Ends Well	1603
26	Shakespeare, Late	Measure for Measure	1604
27	Shakespeare, Late	Othello	1604
28	Shakespeare, Late	King Lear	1605
29	Shakespeare, Late	Macbeth	1606
30	Shakespeare, Late	Antony and Cleopatra	1607
31	Shakespeare, Late	Timon of Athens	1607
32	Shakespeare, Late	The Tragedy of Coriolanus	1608
33	Shakespeare, Late	Cymbeline, King of Britain	1609
34	Shakespeare, Late	The Winter's Tale	1610
35	Shakespeare, Late	The Tempest	1611
36	Shakespeare, Late	Henry VIII	1613

Table 1: The Corpus

Author	# Works	# Words	% of Corpus (in words)
Chapman	11	210,789	10.8
Jonson	14	427,400	22.0
Marlowe	7	121,362	6.2
Middleton	9	215,564	11.1
Shakespeare	36	970,685	50.0

Table 2: Corpus Broken Down by Author

Ben Jonson was a preeminent playwright of his time, and his writing spanned four decades. We obtained 14 works by Jonson totaling over 425,000 words. George Chapman was a prolific dramatist and, like Jonson, a contemporary of Shakespeare, though Chapman is generally better remembered for his translations and poetry. We obtained 11 works by Chapman totaling over 210,000 words. Finally, Thomas Middleton was somewhat younger than the other authors in the corpus, with all of his works appearing after the turn of the seventeenth century, and much of his most famous work published after Shakespeare's 1616 death. Nonetheless, we selected 9 of Middleton's earlier works, totaling some 215,000 words, in order to keep the corpus primarily in the same few decades. Notably, Middleton is thought by many scholars to have at least

collaborated with Shakespeare or adapted/revised Shakespeare's work on several of his later plays, including Macbeth, Timon of Athens, and Measure for Measure (Vickers, 2004). His inclusion allows us to consider what kind of effect editorial changes have on an authorship attribution model. In short, our model results demonstrate that it does not have trouble classifying these works as works of Shakespeare, probably owing to the types of features we selected that are unlikely to be changed by editors.

Our Shakespeare corpus included 36 works from the First Folio. Pericles, Prince of Tyre, was not included because it is a generally accepted collaboration with George Wilkins (Vickers, 2004). Additionally, The Noble Kinsmen was excluded because it is a generally accepted collaboration with John Fletcher. Shakespeare makes up nearly half of our entire corpus with over 970,000 words himself. Marlowe, simply because he had such a small body of work, was the smallest contributor to the corpus, with only 7 works totaling barely 120,000 words.

Dates for works cannot be exact because of the nature of the historical record for this time period, but they were gathered from the Third Edition of the *Annals of English Drama, 975-1700* (Harbage, 1989).

IV. Two Models of Authorship Attribution

As mentioned above, our corpus consists of 77 works (36 attributed to Shakespeare, 7 to Marlowe, 14 to Jonson, 11 to Chapman, and 9 to Middleton). In this set of experiments, there were 77 test trials; in every trial we hold out one work in order to test which author's corpus is the most similar. During training for each trial, we look over the five sets (composing the 76 works *not* being tested) based on the canonical author of each work. Our training gives us five different distributions, one average distribution for each of the canonical authors; the test work's distribution is compared to each of the average distributions to determine which author is most

likely to produce that work. For the work we are testing, we exclude it from the training set of it's canonical author in order to prevent the work from only being classified with it's canonical author simply because we are testing on some of the same language we trained on (hence the only 76 works in training). We train and test twice for each trial using the two different models described below in order to compare their results.

Model 1: General Vocabulary

In the General Vocabulary model, we simply examine all words in the texts. The feature vector (or feature set) of the model is the vocabulary that shows up in the training data of the author plus an "unknown word" that would represent any words that show up in the test document that did not appear in the training data. The value of this unknown word is set to a very small probability to ensure the probability of a word found in the test document is never 0. Finally, for words in the training set that do not occur in the test document, the value of that word's entry in the test work's vector is set to very small probability because probabilities equal to 0 cannot be simply integrated into the metric used to determine the similarity between test document and each training set. We use Kullback–Leibler divergence (Kullback & Leibler, 1951), to calculate the similarities between two distributions (the train and the test, for each of the 5 author's trained on) in order to decide which author is the most likely to write the test work. This operates under the assumption that the most likely author is the one with the most similar average distribution to the test work. In short, it measures how close two points are in an n-dimensional space. It requires two vectors of length n: this would be the size of the training set's vocabulary (number of different types, or words) for each author and the value at each entry in the vector would be the relative frequency of that word (smoothed by very slightly decreasing the probability evenly among all words to allow for the low but nonzero probability of an

unknown word). The relative frequencies of a given word (in n-dimensional space, each word is a dimension) are compared using the KL divergence formula, and the total KL divergence is computed from each test work to each author. The least KL divergence amongst the five author's values corresponds to the closest work/author match and that author is the predicted author of the text.

Note that the General Vocabulary model simply examines the distribution over *all* words in the training set. Contrary to some previous work on the authorship attribution paradigm, as explained in the description of Model 2, this includes content words, such as "sneak" or "Juliet." However, KL divergence is weighted by the frequency of the feature in the training set, so more frequent lexical items contribute more to the KL score. In other words, a big difference in the probability of the word "the" between test and train would have a larger impact on the KL score than a big difference in the probability of the word "sneak" between test and train because the base frequency is much lower. In that way, our General Vocabulary model implicitly takes into account that most function words may be more informative than most content words in telling the difference between authors.

Model 2: Generative Model

In order to follow more closely previous work in the field of authorship attribution, and because we believe there ought to be something more fundamental to the differences between authors than just counting how often each individual word occurs, we developed a generative model that reduced the number of features in our model substantially and, it will be shown, performs at least as well as a model that takes into account the identity and relative frequency of each individual word in it's final calculation of who is the most probable author of the test work.

The features that are exploited in our generative model include function word frequency, frequency of part of speech tags among words that are not on the function word list, and bigram (sequence of two "words," called "two-word collocations" in much of authorship attribution literature including Grieve, 2007) probabilities between these items. We now discuss each feature in turn to motivate it and describe how it was operationalized for our model.

Function words have a long history of consideration in the authorship attribution paradigm. Mosteller and Wallace (1964, 1984), for instance, used distributions of function words to classify sections of the Federalist Papers. They chose this feature set because they suspected that function word distributions are unlikely to fluctuate greatly with the content of the document. One would expect Marlowe to use the word "of" or the word "the" about equally often in *The Jew of Malta* and *Tamburlaine the Great, Part I*, irrespective of the vast difference in the content of these plays. Considering the final results we achieve, it seems clear that we at least do not *lose* power to distinguish between authors by excluding content words from special consideration, once the other features of this model are included. To understand this intuitively, consider that the word "Juliet" may indeed be a good marker for Shakespeare, but knowing that Shakespeare is far more likely to use that word in his corpus than any other author will not help you classify *Hamlet* as being by Shakespeare. Additionally, using distributions over words that *do* vary greatly based on the subject of the work could theoretically lead to two works talking about some character "Edward" being classified as being by the same author because they are both discussing the same topic, or at least appear to be about the same topic because our system uses common strings (series of letters or numbers, etc in text) for frequency counting. While knowing the likelihood an author will use these topic-dependent words may be useful for some kinds of tasks, and could certainly provide some information regarding preferences of word choice or

preferred subjects of different authors, our intention in constructing this model is to build a robust *topic-independent* model that would not give false classifications in an authorship attribution corpus with slightly different characteristics. Additionally, the Generative Model retains greater parsimony because of the decreased feature set size of the Generative Model compared to the General Vocabulary model. Furthermore, Chung & Pennebaker (2007) argue that speakers, and presumably to some similar extent writers, have less cognitive control over their choice of function words, so function word distributions might capture something about the author's deep linguistic tendencies, and it would be harder to emulate an author in order to deceive an authorship attribution model (or, more realistically, a human being reading the text). Finally, many researchers (including Merriam and Matthews, 1994, who worked on the Shakespearean authorship question) have demonstrated that authors do indeed have quite different distributions of frequencies of function words, confirming that this is a promising feature to examine.

Notably, we do not use the small selective lists published by other studies, such as Mosteller and Wallace (1964) or Binongo (2003). Instead we use a larger definition of "function words" to include other kinds of content-independent words. This allows for certain adverbs, such as "respectively," and certain verbs, such as "seemed," and many other words which would not typically be considered function words by linguists or cognitive scientists. Our list totaled 710 words. In order to avoid confusion, we will henceforth refer to our list as a list of "stop words," borrowing the term from computer scientists and natural language processing researchers interested in information retrieval. When trying to determine what a user is searching for and what results would be most helpful, search engines typically filter out stop words to improve efficiency of the algorithms and achieve greater specificity. We, on the other hand, want

to take advantage of these "throw-away" lexical items to capture something about the author's style.

Like function words, parts of speech should not vary substantially with the topic of the text (though between genres – novels and drama, for instance – one may indeed expect differences in both of these feature sets). Intuitively, the probability that an author uses an adverb or adjective in a monologue, or the probability that an author uses proper nouns instead of pronouns in a dialogue could reveal some stylistic tendencies or preferences of the individual. With the dawn of reliable automated statistical natural language parsing systems, it has been possible to examine the frequency distributions of parts of speech between authors to examine whether this syntactic feature is a useful one. Several studies have demonstrated their usefulness in the authorship attribution paradigm (see, e.g. Baayen et al, 1996; Chaski, 2005).

We use the highly accurate Stanford tagger (Toutanova et al, 2003) to obtain the most likely part of speech tag for each word in all the works of the corpus. There are 40 parts of speech tags; they were specified during the creation of the Penn Tree Bank and are used in almost all computational linguistic research in English that involves syntactic parsing (Marcus et al, 1993). It should be noted that the high accuracy of the Stanford parser is known for modern newspaper corpora; while there may not be as high of an accuracy of this tagger on an Elizabethan corpus, this would be a fault of any available pre-trained system.

Typically, bigram probabilities are the relative frequencies of collocations of two words. These features have been examined by many researchers in the authorship attribution paradigm, as well (see, e.g. Merriam, 1979, 1980, 1982; Hoover, 2002). However, in our Generative Model, we are computing bigram probabilities along the feature sets described above. In other words, the bigram "Hell hath" would be converted by our system to "[Noun] [Verb]." The bigram "hath no"

would be converted to "[Verb] no" because "no" is one of the stop words in our list. These bigrams may be able to capture some potentially interesting stylistic cues; for instance, how often does one author use a split infinitive? Similar types of cues have been shown to be useful in studies of short sequences of parts of speech applied to the authorship attribution paradigm (see, e.g. Koppel et al, 2002).

In order to find the most probable author of each test work, instead of directly comparing average frequency distributions of each author to the distribution of the work (a simplification of what KL divergence quantifies), we find the probability that the document was written by a given author and choose the one with the highest probability. Obviously to say "the probability that the set of bigrams in the test play was a random sample drawn from an author's training distribution of bigrams" means the same thing as "the probability that the document was written by that author" is an oversimplification. Nonetheless, this is a common statistical measurement in natural language processing, and it will be demonstrated in the validation and then in the results that this indeed successfully determines the attribution of at least the vast majority of texts for which scholars are confident about authorship (i.e. the works included by Middleton, Jonson, and Chapman). For each bigram in the test play, we find the probability of that bigram occurring in each of the training sets of the authors, and the current probability of the test document given the author (for each author) is multiplied by that corresponding probability of the bigram given the author's training data's distribution. At the end of the document, the probability of each author given the test document's distribution is determined by a simple conversion using Bayes rule (Bayes & Price, 1763); the most likely author directly corresponds to the author for whom the probability of the document given the author is highest.

Validation of Models:

As a simple test to validate the models before expanding to the full experiment with all 5 authors and 77 works in our corpus, we first obtained a machine readable version of *Pericles, Prince of Tyre*, which we did not include in our corpus because it is generally accepted as a collaboration between Shakespeare and George Wilkins, and *The Miseries of Enforced Marriage*, the only existing drama ascribed to *only* Wilkins. We prepared the Wilkins work for training in the same way as the works above, removing all text that was not spoken by the actors. We then prepared *Pericles* for training in the same way, but dividing it up based on acts. It is generally thought that Acts I and II were written primarily by Wilkins and Acts III-V were by Shakespeare (Vickers, 2004). We tested this by training our two authorial distributions only on Wilkins' *Miseries* and Shakespeare's *All's Well that Ends Well* and *Merchant of Venice* (these are two comedies – like *Pericles* and *Miseries* – written around the same time at *Pericles*). We then tested on each of the five acts using first the General Vocabulary model and then the Generative Model. The results are shown in Table 3.

Act of *Pericles*	**Predicted Author: General Vocabulary**	**Predicted Author: Generative Model**	**Canonical Author**
Act I	Wilkins	Wilkins	Wilkins
Act II	Wilkins	Wilkins	Wilkins
Act III	Shakespeare	Shakespeare	Shakespeare
Act IV	Shakespeare	Shakespeare	Shakespeare
Act V	Shakespeare	Shakespeare	Shakespeare
Accuracy:	**100**	**100**	**-**

Table 3: Results of *Pericles, Prince of Tyre* Validation Experiment

In this article, though we are trying to determine who the most likely author of each work is without regard for historical evidence, we report the degree of agreement of each model with the canonical divisions of works by author as "accuracy," or the percentage of works classified to the same author to whom it is typically ascribed. We do this for ease of reporting and

understanding the terminology of the paper. The accuracy for each model above is 100%, predicting based only on the distributions of the training sets specified and the distribution of features within each act exactly the same authors that are historically thought to have penned them. This gives us some confidence that our models are picking out useful and author-specific measures. This is again supported by our highly accurate classification of the non-Shakespearean-candidate authors in the 5-author experiments to follow.

V. The 5-Author Classification Task:

The 5-Author Classification Task was conducted as described before, holding out one work and training on the rest of the corpus. The results of the experiment are shown in Table 4. Plays that were misclassified by at least one model, according to the canonical divisions, are highlighted.

Title of Play	Predicted Author: General Vocabulary	Predicted Author: Generative Model	Canonical Author
The Blind Beggar of Alexandria	Chapman	Chapman	Chapman
An Humorous Day's Mirth	Chapman	Chapman	Chapman
All Fools	Chapman	Chapman	Chapman
May Day	Chapman	Chapman	Chapman
Sir Giles Goosecap	Chapman	Chapman	Chapman
The Gentleman Usher	Chapman	Chapman	Chapman
Bussy D'Ambois	Chapman	Chapman	Chapman
The Widow's Tears	Chapman	Chapman	Chapman
Monsieur D'Olive	Chapman	Chapman	Chapman
Conspiracy and Tragedy of Charles…	Chapman	Chapman	Chapman
The Revenge of Bussy D'Ambois	Chapman	Chapman	Chapman
A Tale of a Tub	Jonson	Jonson	Jonson
The Case is Altered	Shakespeare	Chapman	Jonson
Every Man in His Humour	Jonson	Jonson	Jonson
Cynthia's Revels…	Jonson	Jonson	Jonson
Poetaster	Jonson	Jonson	Jonson
Volpone	Jonson	Jonson	Jonson
The Alchemist	Jonson	Jonson	Jonson
Catiline His Conspiracy	Jonson	Jonson	Jonson
Love Restored	Jonson	Jonson	Jonson

Bartholomew Fair	Jonson	Jonson	Jonson
The Devil is an Ass	Jonson	Jonson	Jonson
The Staple of News	Jonson	Jonson	Jonson
The New Inn, or The Light Heart	Jonson	Jonson	Jonson
The Magnetic Lady	Jonson	Jonson	Jonson
Dido	Marlowe	Marlowe	Marlowe
I Tamburlaine the Great	Marlowe	Marlowe	Marlowe
II Tamburlaine the Great	Marlowe	Marlowe	Marlowe
The Jew of Malta	Shakespeare	Shakespeare	Marlowe
Doctor Faustus	Shakespeare	Shakespeare	Marlowe
Edward II	Shakespeare	Marlowe	Marlowe
The Massacre at Paris	Marlowe	Marlowe	Marlowe
The Family of Love	Jonson	Shakespeare	Middleton
The Phoenix	Middleton	Middleton	Middleton
A Trick to Catch the Old One	Middleton	Middleton	Middleton
The Five Gallants	Middleton	Middleton	Middleton
No Wit, No Help Like a Woman's	Middleton	Middleton	Middleton
The Second Maiden's Tragedy	Middleton	Middleton	Middleton
A Chaste Maid in Cheapside	Middleton	Middleton	Middleton
The Witch	Middleton	Middleton	Middleton
Hengist, King of Kent…	Middleton	Middleton	Middleton
I Henry VI	Shakespeare	Marlowe	Shakespeare
II Henry VI	Shakespeare	Shakespeare	Shakespeare
III Henry VI	Shakespeare	Shakespeare	Shakespeare
The Life and Death of King John	Shakespeare	Shakespeare	Shakespeare
The Comedy of Errors	Shakespeare	Shakespeare	Shakespeare
The Taming of the Shrew	Shakespeare	Shakespeare	Shakespeare
The Tragedy of King Richard III	Shakespeare	Shakespeare	Shakespeare
The Two Gentlemen of Verona	Shakespeare	Shakespeare	Shakespeare
Titus Andronicus	Shakespeare	Shakespeare	Shakespeare
Love's Labour's Lost	Shakespeare	Shakespeare	Shakespeare
Richard II	Shakespeare	Shakespeare	Shakespeare
A Midsummer-Night's Dream	Shakespeare	Shakespeare	Shakespeare
Romeo and Juliet	Shakespeare	Shakespeare	Shakespeare
The Merchant of Venice	Shakespeare	Shakespeare	Shakespeare
I Henry IV	Shakespeare	Shakespeare	Shakespeare
II Henry IV	Shakespeare	Shakespeare	Shakespeare
The Merry Wives of Windsor	Shakespeare	Shakespeare	Shakespeare
Much Ado about Nothing	Shakespeare	Shakespeare	Shakespeare
As You Like It	Shakespeare	Shakespeare	Shakespeare
Henery V	Shakespeare	Shakespeare	Shakespeare
Julius Caesar	Shakespeare	Shakespeare	Shakespeare
Hamlet	Shakespeare	Shakespeare	Shakespeare
Twelfth Night, or What You Will	Shakespeare	Shakespeare	Shakespeare
Troilus and Cressida	Shakespeare	Shakespeare	Shakespeare

All's Well That Ends Well	Shakespeare	Shakespeare	Shakespeare
Measure for Measure	Shakespeare	Shakespeare	Shakespeare
Othello	Shakespeare	Shakespeare	Shakespeare
King Lear	Shakespeare	Shakespeare	Shakespeare
Macbeth	Shakespeare	Shakespeare	Shakespeare
Antony and Cleopatra	Shakespeare	Shakespeare	Shakespeare
Timon of Athens	Shakespeare	Shakespeare	Shakespeare
The Tragedy of Coriolanus	Shakespeare	Shakespeare	Shakespeare
Cymbeline, King of Britain	Shakespeare	Shakespeare	Shakespeare
The Winter's Tale	Shakespeare	Shakespeare	Shakespeare
The Tempest	Shakespeare	Shakespeare	Shakespeare
Henry VIII	Shakespeare	Shakespeare	Shakespeare
Accuracy	**93.5**	**93.5**	**-**

Table 4: Results of 5-Author Classification Task

It is clear from the high accuracy rate of the experiment for the three "non-condidates" (General Vocabulary: Chapman – 100%, Jonson – 92.9%, Middleton – 88.9%; Generative Model: Chapman – 100%, Jonson – 92.9%, Middleton – 88.9%), that both our models are capturing author-specific features. It also appears that any revising or editing Middleton may have done of the works canonically attributed to Shakespeare has not been detected as being more like Middleton than Shakespeare. This is unsurprising if Middleton's involvement was minimal (see also Collins et al, 2004).

It should be noted, though, that there is some apparent "noise" in the data. *The Case is Altered* appears to be more unlike Jonson's works than the rest of his canon. Indeed, we learned after running these experiments that this play's authorship has been brought into question. There are suggestions that the last 2 acts of the 5-act play were not Jonson's, citing a different theatrical style, allusions to works appearing after the supposed date of this play, and a lack of mention of the play by Jonson and failure to include it in early Jonson folios (Loxley, 2002).

The Family of Love also appears to be dissimilar to Middleton's corpus. Again, after the experiment, we discovered that indeed, the authorship of this play has been challenged. It may have been a collaboration between Middleton and Thomas Dekker. More recently, a third

candidate has gained popularity as the author of the work, namely Lording Barry (Taylor et al, 1999).

For both of these works, though, they are not consistently classified by the two models as being by another single author in our corpus, so we have little reason to suspect that one of our other authors had a hand in these disputed works. Instead, it seems that perhaps the true author of these works, if they are indeed not truly by the author to whom they are ascribed, is closer to the authors predicted by the models in his/her frequency distribution than to the canonical author. We did not exclude these works from the corpus following these discoveries because it is certainly the case that some of the correctly classified works are also the subjects of hotly contested authorship questions.

As for the other misclassifications, all are between Marlowe and Shakespeare. Two works canonically attributed to Marlowe, *The Jew of Malta* and *Doctor Faustus* are classified by both models as more similar to the Shakespearean corpus. *Edward II*, also by Marlowe, was classified as a work by Shakespeare only by the General Vocabulary model. *Henry VI, Part I*, a work canonically attributed to Shakespeare and notably the subject of many debates and much scholarship regarding authorship attribution (see, e.g. Merriam, 1998 which suggests Marlowe's hand in the play), was classified by the Generative Model as a work of Marlowe.

However, it is notable that the majority of the works in these author's corpora (General Vocabulary: Shakespeare – 100%, Marlowe – 57.1%; Generative Model: Shakespeare – 97.2%, Marlowe – 71.4%) were classified along the canonical lines. This accuracy is extremely high, overall. It appears clear that our models have "learned the difference" between the corpora we label "Shakespeare" and "Marlowe," but the above listed hiccups may provide interesting clues nonetheless.

Furthermore, one major confound in our study is that the size of Marlowe's corpus is so tiny. In fact, we see that his accuracy rate (General Vocabulary: 57.1%, Generative Model: 71.4%) is the lowest by far of our authors, and his corpus is only 7 works and barely 121,000 words. Perhaps his misclassifications are related to the small training set size. When a work that is tens of thousands of words long (*Doctor Faustus*: 13,883 words; *The Jew of Malta*: 21,544 words; *Edward II*: 25,029 words) is removed from the Marlowe corpus in order to test it, the corpus size is miniscule compared to Shakespeare, who has nearly a million words – about 10 times the size. This could mean that Marlowe's distribution is not "smooth" enough, or averaged over enough words to get good average frequencies of the features. Finally, it should be noted that two works in the canons of two different authors are also misclassified as being most like Shakespeare by one model each. This is notable because it should be remembered that in our experiments, we are more likely to see misclassifications of works to Shakespeare independently of who the canonical author is. This might be because our Shakespearean corpus is the best smoothed. The question is discussed in more detail later.

Despite the misclassifications between the two corpora of special interest (with respect to the canonical corpus divisions), the Marlowe-Shakespeare misclassification rate is low, and not much higher than the misclassification rate of some of our other authors (if we include in our analysis the authorship issues discovered as a result of finding the anomalies). The low Marlowe-Shakespeare misclassification rate is not supportive of the hypothesis that these two authors are the same individual. Examining the clues, however, we see that the plays that are misclassified between the two are "later" Marlowe and the earliest Shakespeare works. The Marlovian Theory would predict that, in fact, Shakespeare was simply "Late Marlowe." Indeed, it may be that our models are so precise that we are not only learning the distinction between authors, but when we

divide a person into an "early" and a "late" version of themselves, the models are learning something about how the style changed with the age of our authors. Our next set of experiments attempted to test this concern, splitting our Shakespeare corpus into an "Early Shakespeare" and a "Late Shakespeare" corpus. We predicted that, if the two corpora were really representative of a shifting style of a single individual over his lifetime, we would see about the same rate of misclassification between Marlowe and Early Shakespeare as between Early Shakespeare and Late Shakespeare. This hypothesis assumes a somewhat linear curve in style shift, which need not be the case. However, we begin with this hypothesis.

VI. Authorship Attribution with Different Ages of the Same Author

The question addressed here is an understudied question in the field of stylometry. To what extent does an author's style drift or shift over time, and can such a change in style be captured by authorship attribution models? While work has been done in forensic linguistics to examine the question of whether one can approximate the age of an individual, these models typically use blogs or emails and features that include emoticons or slang terms that are apparently particular to different age groups (Schler et al, 2006). But our question is slightly different. Could a model that successfully distinguishes between works written by James Madison and Alexander Hamilton also distinguish between works written by James Madison as a young man and works written by James Madison as an older man if his writings are divided into two adjacent decades?

We first divided our Shakespeare corpus into two pieces. We chose to use 1601 as the last year in the "Early Shakespeare" date range because it has been noted by some scholars as a turning point in his work because it is the date after which most of his major tragedies were written. The specific date, however, is if little importance, and the same experiments yielded

approximately comparable results (slightly lower overall accuracy) with a division making the number of works in Early and Late Shakespeare equal.

The Authorship Attribution with Different Ages of the Same Author task was conducted in the exact way that the 5-Author Classification Task was conducted, but instead had 6 authors. One work was held out and the classifiers were trained on the rest of the corpus. The results of the experiment are shown in Table 5. Plays that were misclassified by at least one model, according to the canonical divisions, are highlighted.

Title of Play	Predicted Author: General Vocabulary	Predicted Author: Generative Model	Canonical Author
The Blind Beggar of Alexandria	Chapman	Early Shakespeare	Chapman
An Humorous Day's Mirth	Chapman	Early Shakespeare	Chapman
All Fools	Chapman	Chapman	Chapman
May Day	Chapman	Chapman	Chapman
Sir Giles Goosecap	Chapman	Chapman	Chapman
The Gentleman Usher	Chapman	Chapman	Chapman
Bussy D'Ambois	Chapman	Chapman	Chapman
The Widow's Tears	Chapman	Chapman	Chapman
Monsieur D'Olive	Chapman	Chapman	Chapman
Conspiracy and Tragedy of Charles…	Chapman	Chapman	Chapman
The Revenge of Bussy D'Ambois	Chapman	Chapman	Chapman
A Tale of a Tub	Jonson	Jonson	Jonson
The Case is Altered	Late Shakespeare	Early Shakespeare	Jonson
Every Man in His Humour	Jonson	Jonson	Jonson
Cynthia's Revels…	Jonson	Jonson	Jonson
Poetaster	Jonson	Jonson	Jonson
Volpone	Jonson	Jonson	Jonson
The Alchemist	Jonson	Jonson	Jonson
Catiline His Conspiracy	Jonson	Jonson	Jonson
Love Restored	Jonson	Jonson	Jonson
Bartholomew Fair	Jonson	Jonson	Jonson
The Devil is an Ass	Jonson	Jonson	Jonson
The Staple of News	Jonson	Jonson	Jonson
The New Inn, or The Light Heart	Jonson	Jonson	Jonson
The Magnetic Lady	Jonson	Jonson	Jonson
Dido	Early Shakespeare	Early Shakespeare	Marlowe
I Tamburlaine the Great	Marlowe	Marlowe	Marlowe
II Tamburlaine the Great	Marlowe	Marlowe	Marlowe

The Jew of Malta	Late Shakespeare	Early Shakespeare	Marlowe
Doctor Faustus	Early Shakespeare	Early Shakespeare	Marlowe
Edward II	Early Shakespeare	Early Shakespeare	Marlowe
The Massacre at Paris	Marlowe	Marlowe	Marlowe
The Family of Love	Jonson	Early Shakespeare	Middleton
The Phoenix	Middleton	Middleton	Middleton
A Trick to Catch the Old One	Middleton	Middleton	Middleton
The Five Gallants	Middleton	Middleton	Middleton
No Wit, No Help Like a Woman's	Middleton	Middleton	Middleton
The Second Maiden's Tragedy	Middleton	Middleton	Middleton
A Chaste Maid in Cheapside	Middleton	Middleton	Middleton
The Witch	Middleton	Middleton	Middleton
Hengist, King of Kent…	Middleton	Middleton	Middleton
I Henry VI	Early Shakespeare	Early Shakespeare	Early
II Henry VI	Early Shakespeare	Early Shakespeare	Early
III Henry VI	Early Shakespeare	Early Shakespeare	Early
The Life and Death of King John	Early Shakespeare	Early Shakespeare	Early
The Comedy of Errors	Early Shakespeare	Early Shakespeare	Early
The Taming of the Shrew	Early Shakespeare	Early Shakespeare	Early
The Tragedy of King Richard III	Early Shakespeare	Early Shakespeare	Early
The Two Gentlemen of Verona	Late Shakespeare	Early Shakespeare	Early
Titus Andronicus	Early Shakespeare	Early Shakespeare	Early
Love's Labour's Lost	Early Shakespeare	Early Shakespeare	Early
Richard II	Early Shakespeare	Early Shakespeare	Early
A Midsummer-Night's Dream	Early Shakespeare	Early Shakespeare	Early
Romeo and Juliet	Early Shakespeare	Early Shakespeare	Early
The Merchant of Venice	Early Shakespeare	Early Shakespeare	Early
I Henry IV	Early Shakespeare	Early Shakespeare	Early
II Henry IV	Early Shakespeare	Early Shakespeare	Early
The Merry Wives of Windsor	Late Shakespeare	Early Shakespeare	Early
Much Ado about Nothing	Early Shakespeare	Early Shakespeare	Early
As You Like It	Late Shakespeare	Early Shakespeare	Early
Henery V	Early Shakespeare	Early Shakespeare	Early
Julius Caesar	Late Shakespeare	Early Shakespeare	Early
Hamlet	Late Shakespeare	Late Shakespeare	Early
Twelfth Night, or What You Will	Late Shakespeare	Late Shakespeare	Early
Troilus and Cressida	Late Shakespeare	Late Shakespeare	Late
All's Well That Ends Well	Late Shakespeare	Late Shakespeare	Late
Measure for Measure	Late Shakespeare	Late Shakespeare	Late
Othello	Late Shakespeare	Late Shakespeare	Late
King Lear	Late Shakespeare	Late Shakespeare	Late
Macbeth	Late Shakespeare	Late Shakespeare	Late
Antony and Cleopatra	Late Shakespeare	Late Shakespeare	Late
Timon of Athens	Late Shakespeare	Late Shakespeare	Late
The Tragedy of Coriolanus	Late Shakespeare	Late Shakespeare	Late

Cymbeline, King of Britain	Late Shakespeare	Late Shakespeare	Late
The Winter's Tale	Late Shakespeare	Late Shakespeare	Late
The Tempest	Late Shakespeare	Late Shakespeare	Late
Henry VIII	Late Shakespeare	Late Shakespeare	Late
Accuracy	**84.4**	**87.0**	**-**

Table 5: Results of Authorship Attribution with Different Ages of the Same Author task

The overall accuracy of the models decrease when dividing Early and Late Shakespeare, due, largely, to the 6 misclassifications from the Early to the Late Shakespeare corpus. The last two works written by our Early Shakespeare (*Hamlet* and *Twelfth Night*) appear to be more like Late Shakespeare in both models. For the General Vocabulary model, we see several more Early to Late Shakespeare misclassifications. We see the same noise in the corpus with respect to the two plays of questionable authorship *The Case is Altered* and *The Family of Love*. We are still highly accurate with our classifications of other authors not of special interest to us (General Vocabulary: Chapman – 100%, Jonson – 92.9%, Middleton – 88.9%; Generative Model: Chapman – 81.8%, Jonson – 92.9%, Middleton – 88.9%).

Besides the misclassifications between Early and Late Shakespeare, no works by Shakespeare are misclassified. Notably, no works canonically by Early or Late Shakespeare are classified as being more similar to Marlowe. Now, apparently, *Henry VI, Part I* is more like Early Shakespeare than Marlowe according to both models. The Marlowe works which were misclassified before are still misclassified, mostly to Early Shakespeare, except *The Jew of Malta* which looks more like Late Shakespeare than either Marlowe or Early Shakespeare to the General Vocabulary model. The only other new finding is that, now, the model classify *Dido, Queen of Carthage*, as being more similar to Early Shakespeare than Marlowe. This pushes the classification rate for Marlowe down (General Vocabulary: 42.9%; Generative Model: 42.9%). Of course, recall that there are only 7 works in the Marlowe corpus, so one or two more incorrect

is a huge change in the accuracy rating compared to one or two more Shakespearean works incorrect.

The accuracy of the models remains high. This confirms that our models are generally powerful enough to detect shifts in an author over time, if it is a little rough around the edges as would be expected for a gradual change in authorial style. The misclassification rate between Shakespeare and Marlowe has changed slightly; while it remains low, but not zero, the interesting point is that the misclassification rate is not symmetric. Misclassifications of works by Marlowe to Shakespeare are far more common than the nonexistent misclassifications from Shakespeare to Marlowe. Again, this could be due to the sizes of Marlowe's and Shakespeare's corpus, as explained below.

VII. A Note on the Asymmetry of Our Results

While there is nothing in the mathematics of our comparison techniques that require misclassifications to be symmetric, we started our work assuming they would be. However, in fact, it is clear from our data that, while confusions between the other authors are approximately symmetric (in that they are extremely rare, and inconsistent when they do occur), confusions between Shakespeare and Marlowe are not symmetric. Only in one case do we see a work canonically attributed to Shakespeare classified as more similar to Marlowe. In many more cases, sometimes indeed in over half the Marlowe corpus, we see the works being classified as being by Shakespeare. Clearly this does not mean that Shakespeare did write Marlowe's works, but Marlowe did not in fact write Shakespeare's works. That would be a very strange conclusion to come to for many reasons, but most importantly because our metrics should not be capable of discriminating between these results. There are at least two possible reasons why our results appear somewhat bizarre in this way.

Firstly, the corpora of Shakespeare, Marlowe, and Jonson, for instance, are not equally confusable. This has to do with the size of the corpora. As previously mentioned, when we remove a work that is, for instance, 20,000 words from Marlowe's corpus, we may only train on about 82% of his original corpus. If we remove a work of the same size from Shakespeare's corpus, we retain 98% of his corpus for training. For Jonson, testing on the same work leaves us 95% of his corpus to train our models. For that reason, Marlowe's distribution is not smooth enough, or averaged over enough data to get well enough approximated average frequency distributions over the features considered by the model. For Jonson and Shakespeare, and presumably for the other authors who only have about 10% of their corpus removed based on the same work's testing status, testing any given work does not significantly shift the average distribution around in the multidimensional space. For each of them, there is still a lot of data contributing to that average. However, for Marlowe, the exclusion of a single play has greater potential to drastically change the average frequencies of the features. Marlowe, therefore, is more confusable in general.

Still the question remains – why would a work of Marlowe be more often classified as more similar Shakespeare rather than, for instance, Jonson? One possible answer is somewhat nuanced, and deals with features that are rare. Though the features of our Generative Model are bigrams, we will provide an example with only a single stop-word. Obviously, if a stop-word is rare, than any single bigram including that stop-word is at least as rare if not rarer. Because Shakespeare's corpus has nearly a million words, rare words are simply more likely to occur in his corpus than in any other corpus. Of our 710 stop-words, 13 are used only by Shakespeare (e.g. "whereupon," "namely"). Words that are rare and occur in Shakespeare's corpus also occur sometimes in other authors' corpora. Marlowe, for instance, exclusively shares 5 words with

Shakespeare. In other words, there are 5 words (e.g. “wherever,” “nowhere”) for which both Shakespeare and Marlowe use the words, but none of the other authors do. Note that there are also words with this characteristic, except that they are shared between Shakespeare and another author. 12 such examples (e.g. “formerly,” “really”) are shared exclusively with Jonson, who has a much larger corpus than Marlowe, and therefore you would expect to see more of these rare words.

Because of the way our Generative Model works, rare words that are exclusively shared between one author and another work are highly important in determining a match. For instance, if all instances of the word “wherever” in Marlowe’s corpus occur in a single work, then when that work is held out for testing, Marlowe and every other author except Shakespeare have 0 instances of the word “wherever.” The training distribution of an author who does not have any examples of some of the words in the test document is heavily penalized for this when using our model. Shakespeare, then, has a distinct advantage in attracting works with rare words in that he is more likely to have those rare words because of his corpus size. As for the other authors who exclusively share words, because their corpus is much larger than Marlowe’s, they are more likely to have other examples of those rare words elsewhere in the corpus. Furthermore, the rest of their averages of the features, also because of corpus size, are likely to be more smoothed out, leading to any discrepancies in rare words to be less important for a non-Marlowe author.

This behavior is capable of accounting for, at least in part, the asymmetry in direction of our misclassifications.

VIII. Unsupervised Clustering Experiment

As noted earlier in the article, all of these previous experiments depend on the assumption that the works attributed to each author are at least by the same person, even if two

of the authors are actually the same person, as well. This is significant because it is possible that Marlowe, or some other author, did indeed write a small number of the works attributed to Shakespeare. The point of the purpose of our article, though, is to ask whether all works by Shakespeare and Marlowe are actually by the same individual. We want to let the unlabeled vectors of data for each work cluster together in the way that minimizes the overall differences in the distributions. In this way, we require no train/test splits that involve labeling the training set with the "canonical author."

This is exactly the methodology we pursue in this experiment. We take 6 works from 2 different authors, from that point forward treating the works as if we do not know anything about who wrote them. We then cycle through every possible division of those works into two groups that could be made out of that set of works. Ideally, we could involve all the works of all the authors at once, but the number of combinations grows incredibly fast with the number of groupings and the total number of works. Because of this, we limit the total number of works in this experiment and only compare works thought to be by *two* different authors at one time. To score the "total variance" of each grouping of the works into two groups, we take the sum of the internal variance of each of the two groups of every combination. We calculate the internal variance of a group based on the absolute differences between each element of a feature vector of a work in that group from the mean of that element over all the vectors in the group. Each feature is equally weighted explicitly (unlike in KL divergence), though larger differences are exaggerated more because we square the difference in probabilities between the feature in the work and its average in the grouping. The feature vector is the union of all the stop-words and most probable POS tags of the non-stop-words that show up in each of the works in the experiment. We think that the internal variance of the works that are written by the same author

will be the smallest among the possible groups, and therefore the sum of the variances of the two internal groups in a combination will be minimized when clustered to the most probable author divisions.

To spell out (in pseudocode) the way we calculate the total variance of a grouping:

for each feature:

find the sum of the frequency of this feature in all the works (call this FeatureWorksSum)

for each work:

add to the internal variance: {(the frequency of the feature in this work) -

(FeatureWorksSum/ number of works)}^2

The groupings were then ranked as a function of their total variance. A ranking of 1 indicates that this is the lowest total variance. We then looked for the canonical division in this list, and found its ranking. Because we had to limit the number of works of each author to include, we randomized this list and ran multiple iterations. The results of several different experiments using this same methodology are showing in Table 6. The statistics given are for the rankings of the canonically correct divisions. The lower the number in the last column, the better the model has decided the author of all of the works in line with the canonical divisions.

Authors Compared	# Works	# Iterations	# Combos	Mean	Best	Worst	Median
Marlowe-Shakespeare	12	100	2047	2.05	2	3	2
Marlowe-Jonson	12	100	2047	84.24	1	507	422
Shakespeare-Jonson	12	100	2047	74.25	48	101	69
EarlyShak-LateShak	12	100	2047	825.39	8	1890	1538
Marlowe-EarlyShak	12	100	2047	49.83	1	231	2

Table 6: Results of Unsupervised Clustering Experiments

It is obvious from this table that the division between Marlowe and Shakespeare is quite clear in this unsupervised method that is built to simply look for the groupings of works that minimizes the variance of each of the two author groupings in the differences between frequencies of the stylistic features we selected – stop-words and part of speech tags of non-stop-words. In fact, it appears that Marlowe and Shakespeare have a clearer division than, for instance, Jonson and Shakespeare or Jonson and Marlowe, on average. We have no explanation for why this would be the case, and it is unclear to what extent it could be read as a contradiction of our previous set of experiments because the type of model and comparison metric are so different.

Indeed we also see that the model does an excellent job of distinguishing between works attributed to Marlowe and those canonically by Early Shakespeare. However, the model seems to have difficulty correctly dividing Early and Late Shakespeare. This indicates that this model is finding huge differences between Marlowe and Shakespeare, but cannot find as much of a difference between Early and Late Shakespeare, which would seem to indicate that it is unlikely that Marlowe and Shakespeare are the same person. One interesting difference in this model is that it exploits large differences between feature probabilities equally among frequent and rare items.

IX. Conclusion

We have presented three models and performed different experiments measuring different features that all deal with the Marlowe-Shakespeare authorship debate. What conclusions can we make about the authorship of the examined works from our Elizabethan corpus? In particular, what conclusions are warranted with respect to the oft-posited Shakespeare-Marlowe connection?

First, at the very least, it seems clear that Marlowe's work has a similarity to Shakespeare's that is not shared by other dramatists of their time. Perhaps, as has been suggested often, Marlowe had an influence on Shakespeare. Perhaps, even, Marlowe contributed to or wrote one (or more) of Shakespeare's plays – the most likely being Henry VI, Part 1. It has received the most recognition as "Marlowesque," and it is the only part of the Shakespeare canon that any of our models attributed to Marlowe under any of the settings.

Nevertheless, to the degree one attaches significance to our modeling, it has to suggest that Marlowe and Shakespeare were not the same person. We now highlight what, to us, are the three data points that most forcefully support that position.

First, the overall accuracy of our model's attributions: counting the two different models (General Vocabulary and Generative Model) and the two different test conditions (Shakespeare with and without the Early/Late division) our models had 172 opportunities to misclassify the two. Marlowe and Shakespeare were only confused a total of 14 times.

Second, the misclassifications of Early and Late Shakespeare: if the two were the same person the most natural assumptions would have led to Early Shakespeare being classified as often as Marlowe as Late Shakespeare. This does not happen; all the misclassifications by both models were to Late Shakespeare.

Finally, there are the clustering experiments, whose strengths and weaknesses compliment the leave-one-out experiments. Leave-one-out compared all of the works, while clustering could only look at a few. Conversely, the size of the Shakespeare corpus compared to the others may bias the leave-one-out results. The clustering experiments find, at the very least, Shakespeare being as easily separable from Marlowe as the other comparisons.

We cannot conclude without again emphasizing the limitations of our techniques. Firstly, programs such as ours in no way "understand" differences in style. At most the results suggest that measurement can be done in the absence of appreciation. Secondly, the question of Marlowe "being" Shakespeare is a question of biology, and style (measured statistically or not) is a distant substitute for, e.g., DNA. Lastly, even within the bounds of stylometry, Marlowe is hard to pin down because of the small corpus that exists for him.

That said, our results are best explained by the assumption that Marlowe is not Shakespeare.

X. References

Ackroyd, P. (2005). *Shakespeare: The Biography*. London, UK: Vintage Books.

Argamon, S., Koppel, M., Fine, J., & Shimoni, A. (2003). Gender, genre, and writing style in formal written texts. *Text*, 23 (3), 321–346.

Baayen, H., van Halteren, H., & Tweedie, F.J. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11, 121–131.

Bayes, T & Price, R. (1763). An Essay towards solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M. A. and F. R. S. *Philosophical Transactions of the Royal Society of London*, 53, 370–418.

Bethell, T. (1991). The Case for Oxford. *Atlantic Monthly*, 268 (4), 45–61.

Binongo, J. (2003). Who wrote the 15th Book of Oz? An application of multivariate analysis to authorship attribution. *Chance*, 16:2, 9-17.

Brian, V. (2004). *Shakespeare, Co-Author: A Historical Study of Five Collaborative Plays*. Oxford, UK: Oxford University Press.

Burger, J. & Henderson, J. (2006). An exploration of features for predicting blogger age. In the *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs* (pp. 47–54). New York, NY: ACM Press.

Chaski, C. (2001). Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, 81, 1–65.

Chaski, C. (2005). Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, 4 (1), 1–13.

Chung, C.K. & Pennebaker, J.W. (2007). The psychological function of function words. In K. Fiedler (Ed.), *Social communication: Frontiers of social psychology* (pp. 343–359). New York, NY: Psychology Press.

Collins, J., Kaufer, D., Vlachos, P., Butler, B. & Ishizaki, S. (2004). Detecting Collaborations in Text: Comparing the Authors' Rhetorical Language Choices in *The Federalist Papers*. *Computers and the Humanities*, 38, 15-38.

Craig & Kinney. (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge, UK: Cambridge University Press.

Eliot, S. (1988). *The Problem of The Reign of King Edward III: A Statistical Approach*. Cambridge, UK: Cambridge University Press.

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22 (3), 251–270.

Harbage, A. (1989). *Annals of English Drama, 975-1700 (3rd Edition)*. Wagonheim, S. (Ed). London, UK: Routledge.

Hoover, D. L. (2002). Frequent word sequences and statistical stylistic. *Literary and Linguistic Computing,* 17, 157–80.

Koppel, M., Argamon, S. & Shimoni, A. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17 (4), 401–412.

Koppel, M., Schler, J. & Argamon, S. (2009). Computational Methods in Authorship Attribution. *Journal of the American Society for Information Science and Technology*, 60 (1), 9-26.

Koppel, M., Schler, J. & Zigdon, K. (2005). Determining an author's native language by mining a text for errors. In *Proceedings of KDD 2005* (pp. 624–628), Chicago, IL.

Kullback, S. & Leibler, R.A. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics*, 22 (1), 79–86.

Loxley, J. (2002). *The Complete Critical Guide to Ben Jonson*. London, UK: Routledge.

Marcus, M.P., Santorini, B. & Marcinkiewicz, M. (1993). Building a large annotated cor- pus of English: The Penn Treebank. *Computational Linguistics*, 19(2),313–330.

Mascol, C. (1888a). Curves of pauline and pseudo-pauline style I. *Unitarian Review*, 30, 452–460.

Mascol, C. (1888b). Curves of pauline and pseudo-pauline style II. *Unitarian Review*, 30, 539–546.

Matthews, R. & Merriam, T. (1993). Neural computation in stylometry: An application to the works of Shakespeare and Fletcher. *Literary and Linguistic Computing*, 8 (4), 203–209.

McMichael, G & Glenn, E. M. (1962). *Shakespeare and his Rivals: A Casebook on the Authorship Controversy*. New York, NY: Odyssey Press.

Mendenhall, T. C. (1901). A mechanical solution to a literary problem. *Popular Science Monthly*, 9, 97–110.

Merriam, T. (1996). Marlowe's hand in Edward III revisited. *Literary and Linguistic Computing*, 11 (1), 19–22.

Merriam, T. (1979). What Shakespeare wrote in Henry VIII (Part I). *The Bard*, 2, 81–94.

Merriam, T. (1980). What Shakespeare wrote in Henry VIII (Part II). *The Bard*, 2, 111–118.

Merriam, T. (1982). The authorship of Sir Thomas More. *ALLC Bulletin*, 10, 1–7.

Merriam, T. (1998). Heterogeneous authorship in early Shakespeare and the problem of Henry V. *Literary and Linguistic Computing*, 13, 15–28.

Merriam, T. & Matthews, R. (1994). Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe. *Literary and Linguistic Computing*, 9, 1–6.

Mosteller, F. & Wallace, D. (1964). *Inference and Disputed Authorship. The Federalist (1st Edition)*. Reading, MA: Addison-Wesley.

Mosteller, F. & Wallace, D. (1984). *Applied Bayesian and Classical Inference: The Case of The Federalist Papers*. New York, NY: Springer-Verlag New York Inc.

Nelson, A. H. (2004). Stratford Si! Essex No. *Tennessee Law Review*, 72 (1), 149–69.

Pinksen, D. (2008). *Marlowe's Ghost: The Blacklisting of the Man Who Was Shakespeare*.

Price, D. (2001). *Shakespeare's Unorthodox Biography: New Evidence of an Authorship Problem*. Westport, CT: Greenwood Press.

Schler, J., Koppel, M., Argamon, S. & Pennebaker, J. (2006). Effects of age and gender on blogging. In *Proceedings of the AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs* (pp. 199–205). Menlo Park, CA: AAAI Press.

Schoenbaum, S. (1991). *Shakespeare's Lives*. Oxford, UK: *Oxford University Press.*

Shaw, R. (2007). *Blank Verse: A Guide to its History and Use.* Athens, OH: Ohio University Press.

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60 (3), 538-556.

Taylor, G., Mulholland P. & MacDonald, J. P. (1999). Thomas Middleton, Lording Barry and the Family of Love. *Papers of the Bibliographical Society of America*, 93, 213–41.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003*, 252-259.

Wells, S. W. (1997). *Shakespeare: A Life in Drama*. New York, NY: W. W. Norton & Company.

Williams, C.B. (1975). Mendenhall's Studies of Word-Length Distribution in the Works of Shakespeare and Bacon. *Biometrika*, 62, 207-212.