Attributing the Authorship of the *Henry VI* Plays by Word Adjacency

Santiago Segarra, Mark Eisen, Gabriel Egan, Alejandro Ribeiro

Shakespeare Quarterly, Volume 67, Number 2, Summer 2016, pp. 232-256 (Article)

Published by Johns Hopkins University Press
DOI: https://doi.org/10.1353/shq.2016.0024

Attributing the Authorship of the *Henry VI* Plays by Word Adjacency

Santiago Segarra, Mark Eisen, Gabriel Egan, and Alejandro Ribeiro

Ads on New York's and London's subway systems for training in shorthand writing used to promise that "if u cn rd ths u cn gt a gd job." Shorthand depends on the inherent redundancy of conventional writing, and to quantify this redundancy the founder of information theory, Claude Shannon, devised an experiment.[1] He took a sentence at random from a Raymond Chandler novel and had an assistant guess each letter in turn. The assistant was told when the guess was correct and was told the correct letter when the guess was wrong. After the first few stabs in the dark, the assistant had the benefit of knowing all of the letters prior to the one being guessed. As we might expect, although the first letters of words were frequently wrong—who knows what word Chandler might use next?—once a letter or two were in place the remainder of each word was often easily guessed. Shannon concluded that overall English prose is about 75 percent redundant.

In this context, "redundancy" means predictability: after the letter *t* the letter *h* is much more likely to follow than *x* is, and directly after *q* the appearance of *u* is almost a certainty. The more likely a particular combination, the less information it carries: the *u* after a *q* is almost completely redundant. Shannon developed the mathematics for quantifying the information carried by any message in any coding system, and it works just as well for sequences of words in a sentence as for sequences of letters in a word. Shannon's work allows us to quantify writers' preferences for putting particular words in particular orders or, more generally, for placing them in proximity to one another.

When attributing authors to works of unknown authorship, evidence from favored phrases (particular words in particular orders) and from collocations (particular words appearing near one another) is frequently employed, although just how to isolate these characteristic phrases and collocations is not agreed upon. MacDonald P. Jackson's approach is to take three- and four-word phrases

[1] C. E. Shannon, "Prediction and Entropy of Printed English," *Bell System Technical Journal* 30.1 (1951): 50–64; James Gleick, *The Information: A History, a Theory, a Flood* (New York: Pantheon, 2011), 229–30.

from the text to be attributed and to search for them in various parts of the Literature Online (LION) database, typically confining his searches to the "Drama" section's holdings for plays first performed between 1590 and 1610.[2] Phrases frequently found in other writers' canons Jackson discards, and for the remaining rare phrases he counts the number of occurrences in each canon. After adjusting for the differing sizes of the canons—Shakespeare's is so large that all other things being equal he would get more hits for that reason alone—Jackson looks for any one writer predominating in the hit list. If one writer has disproportionately more hits than the others, Jackson considers this reasonable evidence for that writer being the author of the text to be attributed. Brian Vickers's method is essentially the same except that, rather than running every short phrase in his sample text through the search engine by hand, he relies on plagiarism detection software to find the matches between the sample and the large corpus of solidly attributed works. And instead of searching in LION, he uses a private database of electronic texts compiled by Marcus Dahl.[3]

Rather than searching for relatively rarely occurring phrases, it is possible to automatically search for quite common phrases or to compare the rates of occurrence of various rare or common words. Hugh Craig, John Burrows, and Arthur Kinney have had considerable success with the last of these approaches.[4] Vickers

[2] MacDonald P. Jackson, "Indefinite Articles in *Titus Andronicus*, Peele, and Shakespeare," *Notes and Queries* 243 (1998): 308–10; MacD. P. Jackson, "*Titus Andronicus* and Electronic Databases: A Correction and a Warning," *Notes and Queries* 244.2 (1999): 209–10; MacDonald P. Jackson, *Defining Shakespeare: "Pericles" as Test Case* (Oxford: Oxford UP, 2003); Jayne M. Carroll and MacD. P. Jackson, "Shakespeare, *Arden of Faversham*, and 'Literature Online,'" *Shakespeare Newsletter* 54.1 (2004): 3, 4, 6; MacDonald P. Jackson, "Shakespeare and the Quarrel Scene in *Arden of Faversham*," *Shakespeare Quarterly* 57.3 (2006): 249–93; MacDonald P. Jackson, "The Date and Authorship of Hand D's Contribution to *Sir Thomas More*: Evidence from 'Literature Online,'" *Shakespeare Survey* 59 (2006): 69–78; MacD. P. Jackson, "Compound Adjectives in *Arden of Faversham*," *Notes and Queries* 251.1 (2006): 51–55; MacD. P. Jackson, "The Authorship of *A Lover's Complaint*: A New Approach to the Problem," *Papers of the Bibliographical Society of America* 102.3 (2008): 285–313; MacDonald P. Jackson, "*A Lover's Complaint*, *Cymbeline*, and the Shakespeare Canon: Interpreting Shared Vocabulary," *Modern Language Review* 103.3 (2008): 621–38; MacD. P. Jackson, "New Research on the Dramatic Canon of Thomas Kyd," *Research Opportunities in Medieval and Renaissance Drama* 47 (2008): 107–27; MacDonald P. Jackson, *Determining the Shakespeare Canon: "Arden of Faversham" and "A Lover's Complaint"* (Oxford: Oxford UP, 2014).

[3] Brian Vickers, "Thomas Kyd: Secret Sharer," *Times Literary Supplement* (18 April 2008): 13–15; Brian Vickers, "The Marriage of Philology and Informatics," *British Academy Review* 14 (2009): 41–44; Brian Vickers, "Disintegrated. Did Thomas Middleton Really Adapt *Macbeth*?," *Times Literary Supplement* (28 May 2010): 14–15; Brian Vickers, "Shakespeare and Authorship Studies in the Twenty-First Century," *Shakespeare Quarterly* 62.1 (2011): 106–42; Brian Vickers, "Identifying Shakespeare's Additions to *The Spanish Tragedy* (1602): A New(er) Approach," *Shakespeare* 8.1 (2012): 13–43.

[4] Hugh Craig and Arthur F. Kinney, eds., *Shakespeare, Computers, and the Mystery of Authorship* (Cambridge: Cambridge UP, 2009); Hugh Craig, "Style, Statistics, and New Models

has strongly condemned counting the frequencies of individual words on the grounds that

> words are not independent but interdependent: one word looks for another. A typical noun phrase includes a substantive, a definite or indefinite article, and a modifier, such as an adjective or superlative. . . . Each of these word classes needs the others. To separate them out reduces language to a severely limited lexicon.[5]

One reason that Craig, Burrows, and Kinney count words rather than phrases and collocations is that the process may be computerized by well-defined and publicly declared algorithms rather than relying upon the investigator to perform manual searches, as Jackson does, or depending upon the operation of an unpublished plagiarism-detection algorithm, as Vickers does. Vickers is right that it would be useful to "develop methods that go beyond the lexicon, beyond the atomistic form of analysis that single words offer, into a 'holistic' method that can respect the phenomenon of language as words that a speaker or writer has joined together in unique sequences."[6]

I. Word Adjacency Networks (WANs)

Vickers's observation that in language "one word looks for another" is correctly imprecise: we cannot yet say much more than that about the general principles by which writers combine words. Vickers and Jackson attend to the close proximities of relatively rare words, one to another, while Craig, Burrows, and Kinney count overall frequencies of relatively frequently occurring words. Ideally, for each text we would count the proximity of every word to every other word, to capture the phenomenon of word-clustering at all levels—among rare words and frequent ones—wherever it occurs. The difficulty is not so much in capturing this vast body of data but in representing it in a form that enables meaningful comparisons between texts.

The technique described here is an application to Shakespearean authorship attribution of what are called Markov chains in order to represent Word Adjacency Networks (WANs) for entire author canons (and subsets thereof)

---

of Authorship," *Early Modern Literary Studies* 15.1 (2009–10): 41 paras; John Burrows and Hugh Craig, "Authors and Characters," in "Stylometry and Authorship Attribution," special issue, *English Studies* 93.3: (2012): 292–309; Hugh Craig, "George Chapman, John Davies of Hereford, William Shakespeare, and *A Lover's Complaint*," *Shakespeare Quarterly* 63.2 (2012): 147–74; Hugh Craig and John Burrows, "A Collaboration about a Collaboration: The Authorship of *King Henry VI, Part Three*," in *Collaborative Research in the Digital Humanities: A Volume in Honour of Harold Short, on the Occasion of His 65th Birthday and His Retirement, September 2010*, ed. Marilyn Deegan and Willard McCarty (Farnham, UK: Ashgate, 2012): 27–65.

[5] Vickers, "Shakespeare and Authorship Studies in the Twenty-First Century," 117.

[6] Vickers, "Identifying Shakespeare's Additions to *The Spanish Tragedy* (1602)," 24.

and to use Shannon's mathematics to compare author-WANs to the WANs for particular works to be attributed.[7] We cannot explain in psychological, neurological, or artistic terms just why various words cluster with other words. We can, however, capture for different authors just how far from one another they tend to place each of a large set of preselected words ("target words") in which we are interested. It turns out that the habits of placement—that is, the choices of words and how far from one another they are placed—vary enough from author to author for this to be a reliable test for identifying authorship when it is unknown or disputed.

To illustrate the technique, we use two short play extracts and attend to the occurrences of just four target words: "with," "and," "one," and "in."

> **With one** auspicious **and one** dropping eye,
> **With** mirth in funeral **and with** dirge in marriage,
> **In** equal scale weighing delight **and** dole.
> (Shakespeare, *Hamlet*, 1.2.11–13)

> I wonder then, that of five hundred, four,
> Should all point **with** their fingers **in one** instant
> At **one and** the same man?
> (Dekker, *Satiromastix*, 1.2.242–44)[8]

Starting with the first target word found in the extract from *Hamlet*, "**With**," we look forward to the following five words, "**one** auspicious **and one** dropping," and notice that two of our target words—"**one**" and "**and**"—fall within this window, the former occurring twice. We record this fact in our network (see figure 1) by drawing an arrowheaded line called an "edge" from the node labeled "**with**" to the node labeled "**one**" and by writing "2" on this line (to represent two occurrences of **one** following **with**). We draw another such edge from **with** to **and**, this time recording the edge weight (as we call it) as 1 for the single occurrence of **and**. Then we move to the next occurrence of a target word, which is "**one**" followed by the five-word window "auspicious **and one** dropping eye." To record these occurrences we create an edge from **one** to **and** with a weight of 1 and an edge from **one** back to itself, weighted 1. Our next target word is "**and**" followed by "**one** dropping eye **With** mirth," so we create an edge from **and** to **one** with a weight of 1 and an edge from **and** to **with** with a weight of 1.

[7] Santiago Segarra, Mark Eisen, and Alejandro Ribeiro, "Authorship Attribution through Function Word Adjacency Networks," *Institute of Electrical and Electronics Engineers (IEEE) Transactions on Signal Processing* 63.20 (2015): 5464–78.

[8] Given that textual provenance is not at issue and that these quotations serve only methodological illustration, we have here used modernized versions of early texts: F in the case of *Hamlet* and the 1602 quarto in the case of *Satiromastix*. See section 7 for a discussion of the provenance and processing of electronic texts used in our testing.

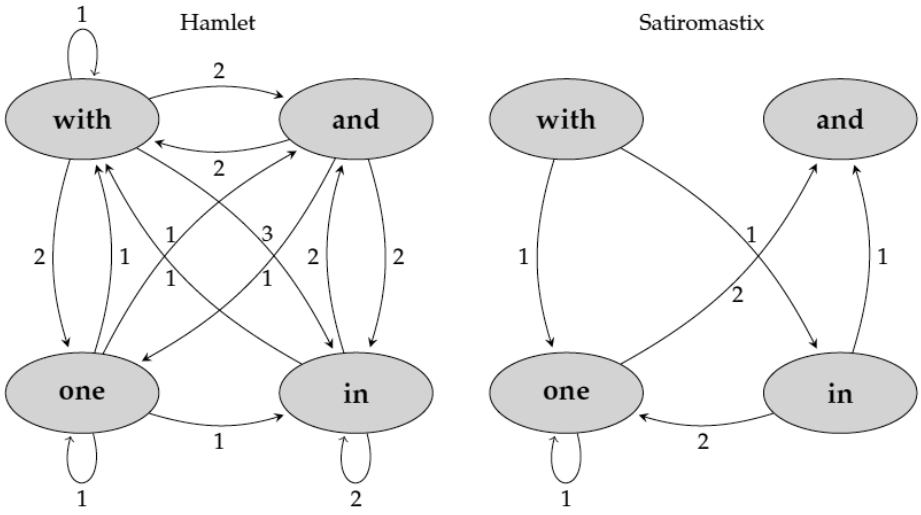

Figure 1. Unnormalized WANs for the extracts from *Hamlet* and *Satiromastix*.

The next target word is "**one**" followed by "dropping eye **With** mirth **in,**" which creates an edge from **one** to **with** (weight 1) and an edge from **one** to **in** (weight 1). Our next target word is "**With**" followed by "mirth **in** funeral **and with,**" which we record as an edge from **with** to **in** (weight 1), and since there already is an edge from **with** to **and**, we do not create a new edge but instead raise the existing edge's weight by one, taking it from 1 to 2. We then create an edge from **with** back to itself (weight 1). The next target word is "**in**" followed by "funeral **and with** dirge **in,**" recorded as an edge from **in** to **and** (weight 1), an edge from **in** to **with** (weight 1), and an edge from **in** back to itself (weight 1). Our next target word is "**and**" followed by "**with** dirge **in** marriage **In.**" Since there already is an edge from **and** to **with**, we raise this existing edge's weight by one, taking it from 1 to 2, and the two occurrences of "**in**" in this window create an edge from **and** to **in** that we give a weight of 2. The next target word is "**with**" followed by "dirge **in** marriage **In** equal," requiring us to increase the weight on the edge from **with** to **in** by two. Our next target word is "**in**" followed by "marriage **In** equal scale weighing," which requires us to raise the weight on the edge from **in** leading back to itself by one. The penultimate target word occurrence in this extract is "**in**" followed by "equal scale weighing delight **and,**" which we record by raising the weight on the edge from **in** to **and** by one. The last occurrence of a target word is "**and**" followed by the single word "dole," which gives us nothing to record.

We repeat this process for the extract from *Satiromastix* to produce its WAN (also figure 1). We can see immediately that this WAN is much less busy than the one for the *Hamlet* extract—fewer edges, lower weights—but it is hard to say how much of the difference is caused by the difference in total occurrences

of the four words we looked for, there being eleven in the *Hamlet* extract and five in the *Satiromastix* extract. Without further processing we can say that in the *Satiromastix* extract the words "**with**" and "**and**" never occur within five words of one another (since no edges connect them) and that "**one**" is our only target word that is followed by another occurrence of itself, recorded as an edge from **one** back to **one**.

The two extracts are the same length, twenty-three words, so they had what we might call an equal opportunity to contain collocations of the four target words. This notion of opportunity is metaphorical: we treat the text as if we did not know which word comes next, playing a version of Shannon's prediction game using early modern plays instead of a Raymond Chandler novel. If someone told us in advance that of the words "**and**," "**one**," and "**in**," Shakespeare would be most likely to use "**in**" shortly after he used "**with**," then on the basis of the *Hamlet* WAN we could say that this extract meets that expectation, for the preference is encoded by the weight of 3 (the highest in this WAN) on the edge from **with** to **in**. Reversing the analogy, we could say that the WAN embodies an expectation of Shakespeare's word choices in an unseen piece of his writing. Naturally, our sample of Shakespeare and our pool of target words would have to be much larger for such an expectation to be realistic.

Although the two extracts each have twenty-three words, the greater occurrence of the target words in the Shakespeare extract prevents us from comparing the raw weights of the WANs' edges. What we really want to know are the proportions of each word choice being made rather than their absolute values. That is, we want to express that each word choice was made not as a raw count but as an act of selection from a range of possibilities. Such an expression is the mathematical way of turning "it was the choice made five times" into "it was the choice made one-time-in-seven when the possibility arose." This is a process called "normalization," and we perform it by dividing the weight of each edge emerging from a node (expressing a particular choice) by the total of all the weights of all the edges (representing all the possible choices) emerging from that node. This tells us for each word the relative frequency with which it is followed, within five words, by each of the other words that are indeed found within five words of it.

In our illustration, we scored "1" if a target word falls anywhere within our five-word window, but we ought to weigh each collocation according to the distance between the words. We do this with a diminishing scale of weights to achieve "proximity scoring" in which, for each match, the farther apart the words are the less their collocation counts in our final summation.[9] Figure 2 shows the

[9] To achieve this, we score 1 (in fact 0.75 to the power of 0) for an occurrence in the first (leftmost) position of the five-word window, 0.75 (that is, 0.75 to the power of 1) for an occurrence

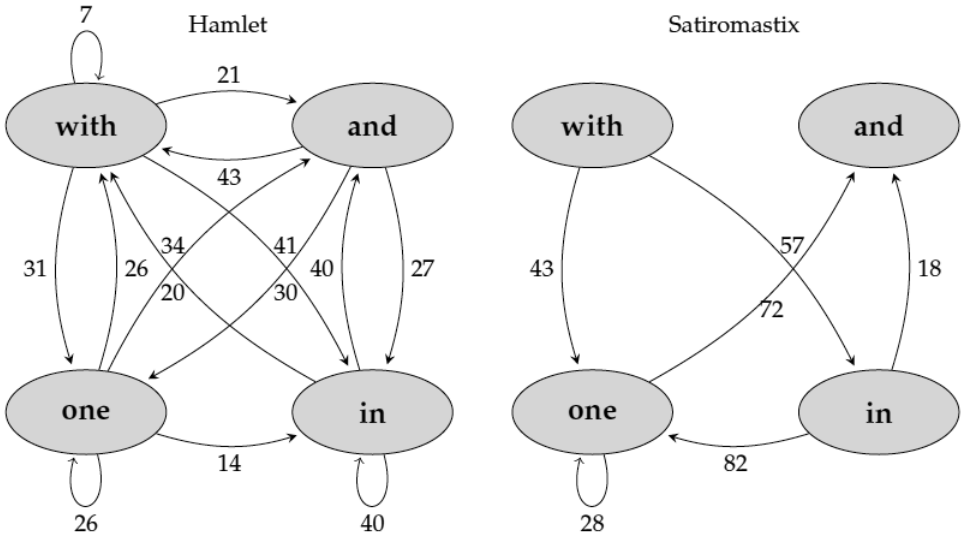

Figure 2. Normalized WANs using proximity scoring and weights scaled to percentages for the extracts from *Hamlet* and *Satiromastix*.

WANs for the extracts from *Hamlet* and *Satiromastix* after we use proximity scoring and then apply normalization, with edge weights presented as percentages—so the weights of edges emanating from each node sum to one hundred—rounded up to whole numbers.

The four target words considered so far are very common in everyday language. As we consider larger sets of target words, however, we will begin to include words that occur considerably less often. When comparing the networks of two texts, the difference between their respective usage of the word "and" should matter more to us than their respective usage of the word "beneath," simply because the word "and" appears more often in English writing. How to factor this in? With edges pointing into and out of each node, the challenge is analogous to calculating the importance of a website by counting (as

in the second position, 0.5625 (0.75 to the power of 2) for the third position, 0.421875 (0.75 to the power of 3) for the fourth position, and 0.31640625 (0.75 to the power of 4) for the fifth position. In general, using a constant close to 1 instead of 0.75 would mean that we do not consider the distance between the target words to matter very much (because when raised to each of the first five powers its gets small quite slowly), while choosing a constant close to 0 means that we consider this distance highly significant (when raised to each of the first five powers it gets small quite rapidly). Experiments on the relationship between this constant and the attribution accuracy of our method—detailed in the article cited in n. 7 above—showed that accuracy is maximized when the constant is between 0.7 and 0.8, hence our selection of the value 0.75 here.

Google's PageRank algorithm does) the links to it from other websites. Thus we rank each node using the ranks of the other nodes that send edges into it.

This recursive approach of defining each rank in terms of all of the other ranks seems to create a chicken-and-egg problem, but in mathematics it is solved by imagining someone traversing the WAN by hopping from node to node for all eternity and choosing which edge to follow for each hop by using the weights as probabilities. The higher the weight, the more often she makes that hop. The longer this continues, the more that the proportion of the hopper's time spent at each node will converge upon the mathematically calculable limit probability, which is the proportion of time spent at each node if the hopping went on forever. These limit probabilities are used in the calculation to measure WANs' likeness by weighting more heavily the results for more commonly used words.

II. Comparing WANs for Different Plays and Writers

We are now in a position to start comparing WANs directly, and for this we want a measure that embodies the totality of the differences between all of the corresponding edges in two WANs. This is calculated using Shannon's mathematics for relative entropy.[10] Shannon's work inaugurated the field of information theory by quantifying the very characteristic we are interested in: the likelihood of a data source (in this case, a writer) emitting a given symbol (call it *y*) immediately or shortly after emitting another symbol (call it *x*). Where the symbols are letters of the alphabet, we already know from experience that certain pairs of symbols are more likely to occur together than others. The letter *t* followed immediately by *h* is such a common combination in English that it has already occurred 396 times in this essay thus far, while *t* followed by *c* is much rarer, having occurred without conscious intention just twice so far, in "match" and "matches." Entropy is a measure of such differential predictability in any symbolic system. Where the symbols are whole words rather than letters, the collocation habits vary considerably from writer to writer, but the way to measure them is the same, employing the mathematics of relative entropy developed by Shannon.

[10] For the technically interested, the procedure is described in G. Kesidis and J. Walrand, "Relative Entropy between Markov Transition Rate Matrices," *Institute of Electrical and Electronics Engineers (IEEE) Transactions on Information Theory* 39.3 (1993): 1056–57. For each edge common to the two WANs, we subtract from the natural logarithm of the weight in the first WAN the natural logarithm of the weight in the second WAN and then multiply this difference by the weight of the edge in the first WAN and then by the limit probability of the node from which this edge originates. This calculation is made for each edge and the values summed to express the total difference. Naturally, it matters which WAN we designate as "first" and which "second," so the procedure is performed twice, switching the designation the second time.

Table 1. Relative entropies (*centinats*) between *2 Henry IV*, *The Tempest*, *Every Man in His Humour*, and *Cynthia's Revels*, omitting ("no comp[arison]") the testing of a play against itself.[a]

	2nd WAN			
1st WAN	*2 Henry IV*	*The Tempest*	*Every Man In*	*Cynthia's Revels*
2 Henry IV	no comp.	4.6	13.8	10.8
The Tempest	4.5	no comp.	19.9	9.1
Every Man In	14.5	20.6	no comp.	6.4
Cynthia's Revels	11.2	9.4	6.4	no comp.

[a] The two Shakespeare plays are closer to one another than they are to the Jonson plays, which are also closer to one another than they are to the Shakespeare plays.

Using the target words "**the**," "**to**," and "**and**," we built word adjacency networks for Shakespeare's *2 Henry IV* and *The Tempest* and Ben Jonson's *Every Man in His Humour* and *Cynthia's Revels*. We then calculated the relative entropies between them—measured in the unit of entropy, *centinats* (*cns*)—shown in table 1. Because in the calculation it matters which WAN we designate as "first" and which "second," we perform each comparison twice, switching the order each time. Thus, each play appears twice in the table: once in the leftmost column when it was the "first" and the other was "second" and once in the top-most row when it was "second" and the other was "first." There is no point comparing each play with itself, so the abbreviation "no comp." for "no comparison" appears where appropriate in table 1.

The relative entropies between *Every Man in His Humour* and *Cynthia's Revels* are 6.4 *cn* for both orderings. These entropies are much smaller than the relative entropies between these two plays and the Shakespeare plays, which range from 9.1 *cn* to 20.6 *cn*. Conversely, the relative entropies between *2 Henry IV* and *The Tempest* are 4.5 *cn* and 4.6 *cn*, which are much smaller than the relative entropies for pairs of plays by different authors. We can use this kind of table as an attribution tool by finding the authorial WAN that is least different from the WAN of the work to be attributed. Where the aim is to exclude candidate authors, large differences in the relative entropies give large degrees of confidence in the exclusion.

In these illustrations of the method, we are still creating WANs that track just four target words. Even with so limited a sample of style, we have WANs that enable us to distinguish our two authors' works with remarkable reliability. In our full analysis we use networks that involve between fifty-five and one hundred words, resulting in more nuanced modeling of writing styles and giving greater numerical separation between texts by different authors.

III. Choosing the Target Words

In general, the more target words we use in a WAN, the more accurate the method. We use so-called "function words," which are words expressing grammatical relationships between other words while carrying little or no lexical value of their own. The role of function words is to bring together the nouns, verbs, and adjectives in order to give a sentence its foundational structure. Typical function words in the English language are prepositions, conjunctions, articles, particles, auxiliary verbs, and pronouns, although linguists differ on which particular examples have so little lexical value as to properly belong to the category.[11] A full list of the function words used in our experiments is given in Appendix 1(a). Words such as nouns and verbs carrying lexical weight might be chosen by an author to suit the particular topic of a play and hence might vary more by genre than by authorship. The choice, the frequency, and (crucially for our method) the relative placing of function words, on the other hand, appear to be an unconscious set of preferences specific to an author.

We noted above that in general the greater the overall frequency of a target word the more weight we should give its role within a WAN, and we derived the appropriate weighting from the same mathematics as Google's PageRank algorithm, which treats an analogous problem. As well as target-word weighting, we must address the problem of target-word selection. We can do better than just using all of our function words as targets in each experiment, since certain words are more discriminating than others for particular cases, and we can identify them. Take Shakespeare and Jonson. First, we rank the function words listed in Appendix 1(a) in order of their frequency of occurrence in the Shakespeare and Jonson canons. We construct an authorial profile WAN (based on the whole authorial canon) and another for each text in each canon using only the two most common words: "and" and "the." We attribute (as if its authorship were unknown) one Jonson play that we held aside (did not use) when making the Jonson profile WAN, and we see how accurate the attribution is, based on just these first two target words.

Say we find that half of the plays (50 percent) are correctly attributed between the two authors. We then add the next most common word from Appendix 1(a)—"to"—to our list of target words, recreate the profile WANs, and retest the attribution of each of the plays. Now perhaps 60 percent of the

[11] The problem of choosing function words is thoroughly explored in Alexis Antonia, "Anonymity, Individuality and Commonality in Writing in British Periodicals, 1830 to 1890: A Computational Stylistics Approach" (PhD diss., University of Newcastle, New South Wales, Australia, 2009), 57–69. Experiments show that our method is not highly sensitive to the small amounts of lexical weight carried by some function words that Antonia would prefer to exclude.

plays are correctly attributed to either Jonson or Shakespeare, so we conclude that adding "to" improved our method. If it had not improved the accuracy, we would exclude it from our list of most discriminating words for this particular comparison. The process is repeated again for networks using the four most common function words, taking note of the total attribution accuracy. We continue increasing the number of target words and freshly attributing each of the plays of known authorship until we have used all 211 target words. The set of target words that produced the greatest accuracy in attributing the plays known to be by Shakespeare or Jonson is then considered the best set for distinguishing between them when we want to attribute plays of unknown authorship for which Shakespeare and Jonson are possible candidates.

IV. Validation

Before applying our tool to some of the most contentious current claims in Shakespearean authorship, we will demonstrate its effectiveness by using it to attribute, as if unknown, the authorship of a set of plays from six of the period's most prominent playwrights. While here we only briefly present the outline and results of the validation process, we refer readers to a separate paper in which the validation process and results for plays from the early modern English period are discussed in greater detail.[12] Our main source of authorship attribution information is the online Database of Early English Playbooks (DEEP) created by Zachary Lesser and Alan B. Farmer and hosted by the University of Pennsylvania—itself based, in part, on the scholarship of Alfred Harbage—and we supplement this with more recent scholarship, such as editions of complete works, where it represents a reasonable consensus.[13] Thus, Appendix 2(a) represents what we believe to be the noncontentious, well-attributed, sole-authored plays of six dramatists from Shakespeare's time. We exclude noncommercial drama, such as court masques and civic entertainments, and plays surviving only as small fragments, such as Jonson's *Mortimer's Fall*. For this validation process, we build a profile WAN for each of the six dramatists using the plays listed.

[12] M. Eisen, S. Segarra, G. Egan, and A. Ribeiro, "Stylometric Analysis of Early Modern Period English Plays," arXiv preprint, arXiv:1610.05670, 2016.

[13] Alfred Harbage, rev. by S. Schoenbaum, *Annals of English Drama 975–1700: An Analytical Record of All Plays, Extant or Lost, Chronologically Arranged and Indexed by Authors, Titles, Dramatic Companies, Etc.* (Philadelphia: U of Pennsylvania P, 1964). For the more recent scholarship see, Gary Taylor and John Lavagnino, gen. eds., *Thomas Middleton and Early Modern Textual Culture: A Companion to the Collected Works* (Oxford: Clarendon Press, 2007); and David Bevington, Martin Butler, and Ian Donaldson, eds., *The Works of Ben Jonson*, 7 vols. (Cambridge: Cambridge UP, 2012).

We attribute the sole-authored plays by computing the relative entropy between each play in Appendix 2(a) and each of the six author profiles, and then we test known collaborative plays against these profiles. Each play is attributed to the author-profile achieving the lowest relative entropy, based on the adjacencies of the one hundred function words listed in Appendix 1(b) that were found in training (based on the full, undisputed sole-authored canon of each dramatist) to be the most discriminating. In order to see the distinctions more clearly, we first calculate for each play its relative entropy with the entire set of all the plays by all six dramatists in order to get a kind of background reading of just how far this particular play differs from the collective norm. Each time that play's relative entropy with the canon of one of the six dramatists is calculated, we deduct from that relative entropy the background reading for that play. It is important to note that no information is lost in this deduction—the relative differences between each play and the author profiles do not change—and it serves only to shift (by the same amount) all of the points on the resulting graph in order to better reveal their relationships. As a result, a final reading of 0 *cn* for the relative entropy between a play and one of the author profiles means that this play is no more or less like that author's work than it is like the combined body of the work of all six authors. A positive relative entropy means that the play is less similar to that author's work than it is to the combined body of all six authors' works, and a negative relative entropy means that the play is more similar to that author's work than it is to the combined body of all six authors' works. Naturally, when a play is being tested against its known author's profile, that play is first taken out of the list of the author's plays, and his profile is created afresh without that play before we compare that play to the author's profile.

Of the 154 plays we considered, we failed to correctly attribute (according to the current consensus) sixteen plays, which is an accuracy of 89.6 percent. In this computation, we consider a collaborative play to be correctly attributed if the best-ranked author is one of the contributors to the play. If we consider the ninety-four plays listed in Appendix 2(a) whose authorship is not in dispute, then the attribution among the six sole-author profiles entails six errors, yielding an overall attribution accuracy of 93.6 percent. These are reasonably high success rates for this kind of attribution method. In addition to attributing full plays, we will analyze the attribution of individual acts and scenes. All authorship attribution methods become less reliable as the sample sizes decrease. To improve accuracy with short texts, we retrain our networks for each test, choosing afresh the most discriminating function words for the question at hand. Networks of seventy-six and fifty-five words were found to produce the best accuracy for the attribution of acts and scenes, respectively (see Appendix 1[c]

and [d]). Our method's accuracy in attributing whole acts of the undisputed plays in Appendix 2(a) is 93.4 percent, about the same as for whole plays.

With individual scenes the accuracy falls off sharply and our method is best used only where the question can be posed as a simple binary choice between two authors, based on strong prior suspicions about authorship. In these cases, the true author may, of course, be someone else not tested for. In such binary tests for individual scenes, our accuracy is 91.5 percent, comparable to that for acts and whole plays. We trust this scene-by-scene approach only where existing scholarship gives good reason to consider just two candidates.

V. The Authorship of *1*, *2*, *and 3 Henry VI*

With our method validated for a great number of plays, we may turn now to the most interesting of the currently contested authorship problems in Shakespeare scholarship: the three parts of *Henry VI* that appeared in the 1623 First Folio. In addition to the six dramatists used in the validation process, we create profiles for two other candidates: Robert Greene and George Peele (see Appendix 2[b]). First we analyze *1 Henry VI* by acts and measure their likeness to the profiles of Shakespeare, Fletcher, Jonson, Marlowe, Middleton, Chapman, Peele, and Greene (see figure 3). Naturally, some of these men are most unlikely candidates for authorship of the play: Fletcher was still in his early teens in the early 1590s and Jonson's earliest known plays did not appear until the end of the 1590s.

Figure 3 suggests that Shakespeare did not write the first act of *1 Henry VI* since it is no more like his profile than it is like Jonson's profile, and Jonson is an implausible candidate. This suggests that it is by someone we did not profile, and Thomas Nashe would be the obvious first choice. Unfortunately, Nashe left just one sole-authored play, *Summer's Last Will and Testament*, which is too small a canon from which to construct a usable author profile. The rest of *1 Henry VI* is, with varying degrees of confidence, assigned to Shakespeare with Marlowe the next closest match for Acts 2 and 4, Greene for Act 3, and Chapman for Act 5. Chapman's earliest known play is *The Blind Beggar of Alexandria*, first performed in 1596.

Because the act-wise analysis suggests Marlowe as a candidate for coauthorship of *1 Henry VI*, we may turn to the scene-wise approach and for each scene ask the binary question "Is this scene more like Shakespeare or Marlowe?" The results are shown in figure 4. For easier visualization, we utilize bar plots for attributing scenes between two authors. We plot the similarity (rather than the difference) between each scene's entropy and each author's profile, so that, for example, a bar of 7 pointing upward means Shakespeare's profile is 7 *cn* closer to the scene than Marlowe's profile is. Thus, the longer the bar for a scene the

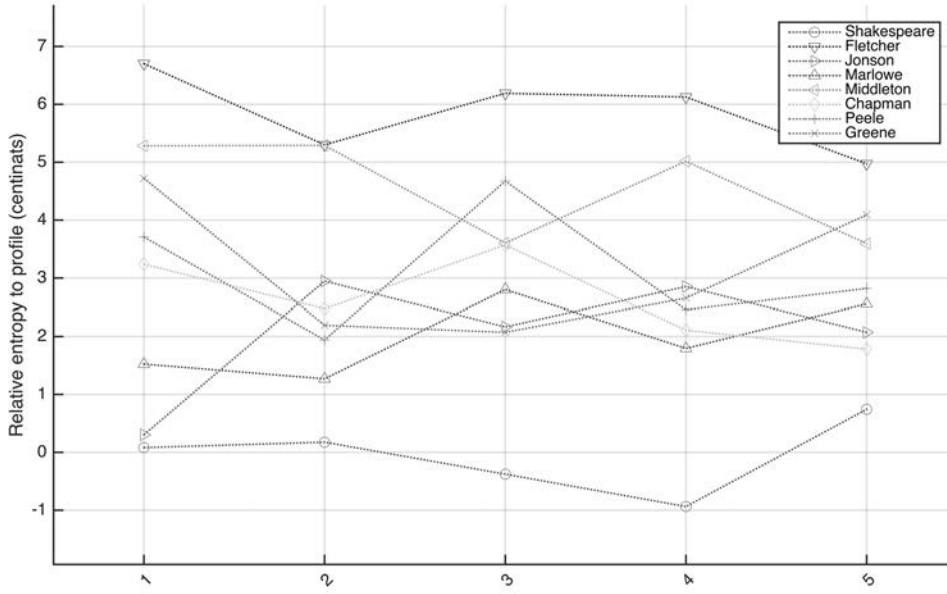

Figure 3. Attribution of individual acts of *1 Henry VI*.

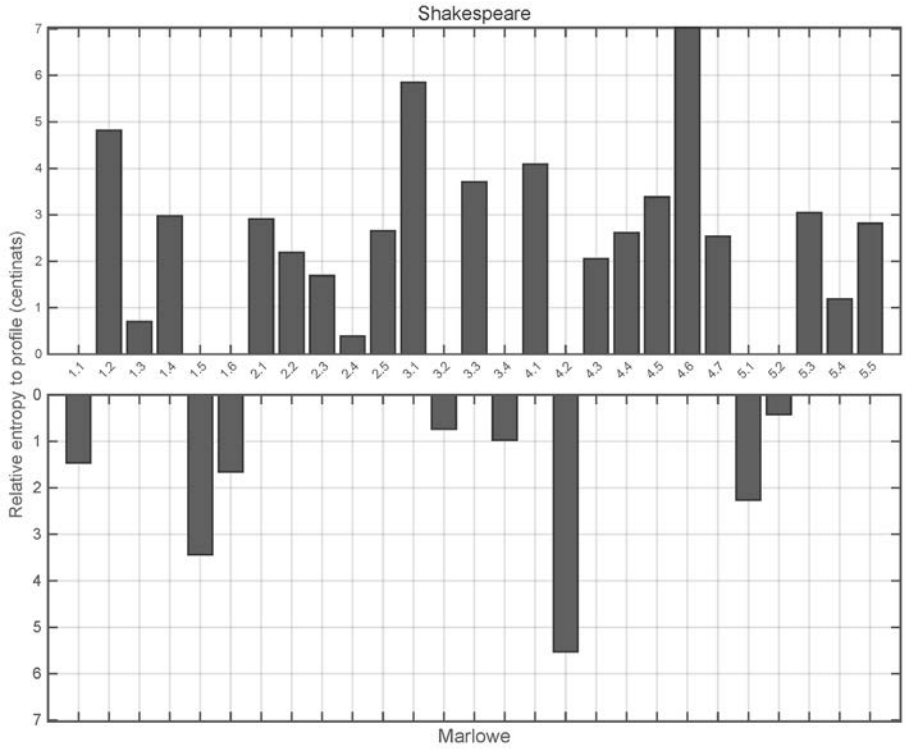

Figure 4. Attribution of individual scenes of *1 Henry VI*.

more that scene is like one author's style rather than the other's style. Corroborating the act-wise analysis, three scenes in Act 1 (1.1, 1.5, and 1.6) go to Marlowe rather than to Shakespeare, as do scenes 3.2, 3.4, 4.2, 5.1, and 5.2. Act 4, scene 2 in particular is attributed to Marlowe by a large margin of almost 6 *cn*. These results broadly agree with Hugh Craig's determination, using an entirely different method, that Marlowe's writing is present in the "middle" and "late" scenes featuring Joan of Arc: 3.2.1–114, 3.3, 4.7, 5.2, 5.3.1–44, and 5.4.[14] Craig grouped 1.2.22–150, 1.5, 1.6, and 2.1 as "early" Joan scenes and found that they collectively test as slightly more Shakespearean than Marlovian. We test these scenes individually and find 1.2 and 2.1 distinctly Shakespearean and 1.5 and 1.6 distinctly Marlovian.

Turning to *2 Henry VI* we repeat the process, looking first at whole acts and attributing them among our eight candidate authors and then using a binary test on each scene. As can be seen in figure 5, we give Act 1 to Marlowe and Acts 2, 3, 4, and 5 to Shakespeare, with Marlowe being a very close second in Act 4. Testing each scene for its likeness to Marlowe and Shakespeare, Marlowe gets 1.1, 1.3, 1.4, 2.2, 2.3, 2.4, 4.1, 4.3, 4.5, and 4.7, and Shakespeare gets 1.2, 2.1, 3.1, 3.2, 3.3, 4.2, 4.4, 4.6, 4.8, 4.9, 4.10, 5.1, 5.2, and 5.3 (see figure 6). The attributions are made with varying degrees of confidence, and the run from 4.3 to 4.8 is really too close to call either way, as shown in figure 6, with both authors' bars being short for these scenes. (We express our confidence here in words rather than in numbers because many more experiments would have to be performed before we could put a value on the confidence levels for closely grouped candidates, and that confidence level would itself have to be qualified with its own uncertainty value.) The 528-line segment from 4.3 to 4.9 may be considered a self-contained, contiguous contribution portraying virtually all of the Jack Cade rebellion. Craig's use of different units of writing makes direct comparison impossible, but his conclusion that from 4.2.160 to 5.1.13 the play is Marlovian is neither confirmed nor denied by our analysis: except for Act 4, scene 10, our method cannot clearly distinguish Shakespeare and Marlowe in this part of the play.[15]

Our act-wise and scene-wise results for *3 Henry VI* are shown in figures 7 and 8. We give Act 1 to Marlowe and Acts 2, 3, 4, and 5 to Shakespeare. By scenes it is 1.1, 2.3, 2.4, 3.3, 4.2, 4.5, 4.7, 5.2, and 5.7 to Marlowe and 1.2, 1.3, 1.4, 2.5, 2.6, 3.1, 3.2, 4.3, 4.4, 4.6, 5.4, 5.5, and 5.6 to Shakespeare with the remaining scenes (2.1, 2.2, 4.1, 4.8, 5.1, and 5.3) too close to call. This case illustrates how important is the (usually unknown) division of authorial labor for

[14] Hugh Craig, "The Three Parts of *Henry VI*," in *Shakespeare, Computers, and the Mystery of Authorship*, 40–77, 61–63 (see n. 4).

[15] Craig, "Three Parts of *Henry VI*," 68–73.

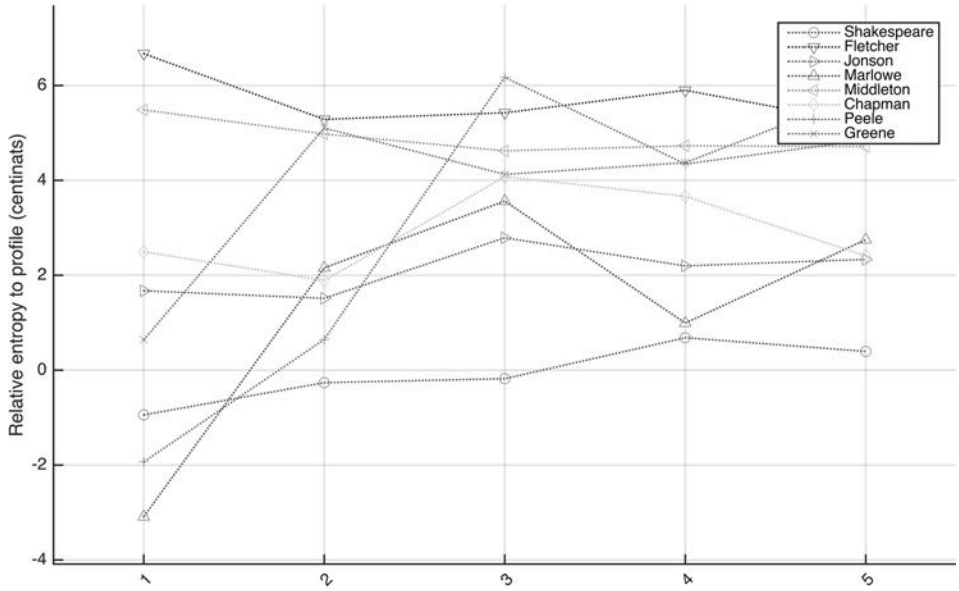

Figure 5. Attribution of individual acts of *2 Henry VI*.

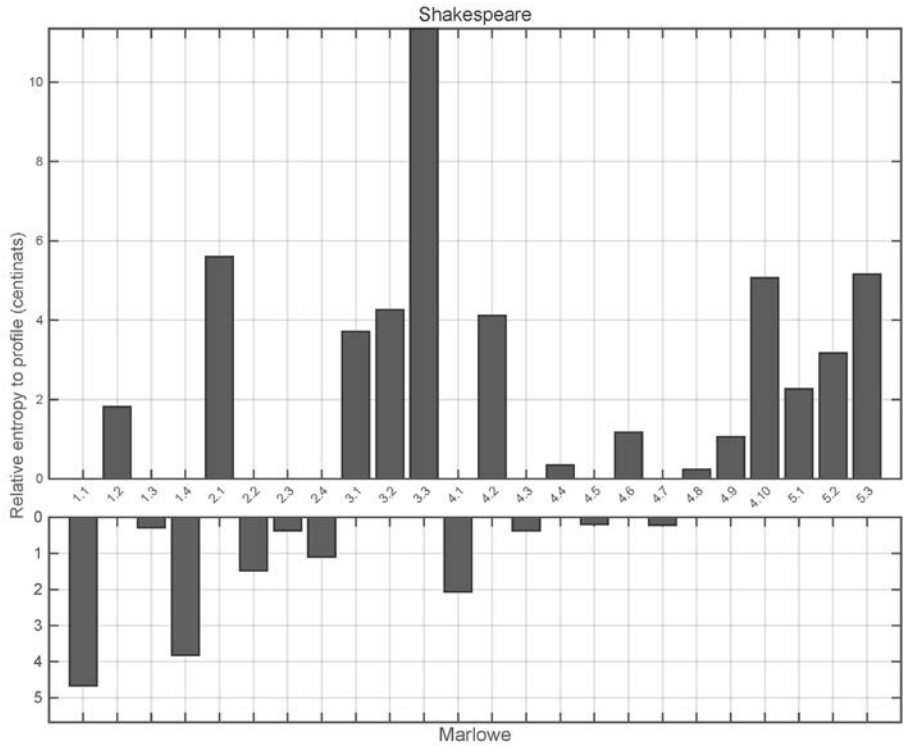

Figure 6. Attribution of individual scenes of *2 Henry VI*.

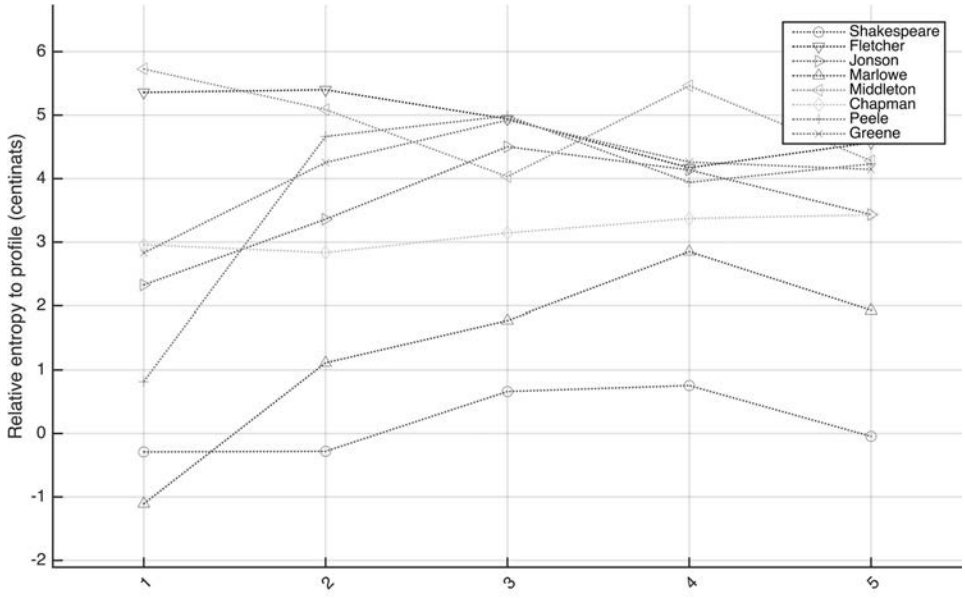

Figure 7. Attribution of individual acts of 3 *Henry VI*.

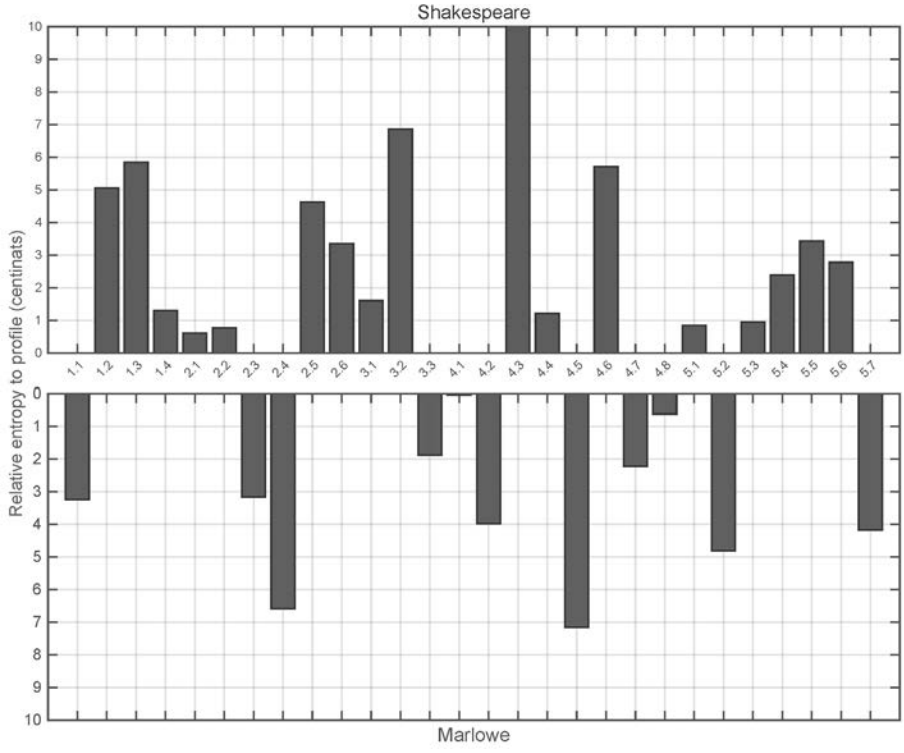

Figure 8. Attribution of individual scenes of 3 *Henry VI*.

such analyses, since the large margin by which Act 4, scene 3 is given to Shakespeare over Marlowe makes the whole of Act 4 seem Shakespearean (as the act-wise analysis finds it) even though it contains scenes (4.2 and 4.5 especially) that are clearly Marlovian.

With the exception of scenes 1.2, 2.4, 4.3, 4.4, 4.6, and 5.7, our results are consistent with those of Craig and Burrows.[16] They find 1.1 to be non-Shakespearean and we agree. They find 1.2 to be non-Shakespearean, and we disagree. They find 1.3, 1.4, 2.1, and 2.2 to be Shakespearean, and we either agree or cannot tell. They find 2.3 to be non-Shakespearean, and we agree. They find 2.4, 2.5, 2.6, 3.1, and 3.2 to be Shakespearean, and we agree except in the case of 2.4 (we say Marlowe). They find 3.3 to be non-Shakespearean, and we agree. They find 4.1 to be Shakespearean, and we cannot tell. They find 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, and 4.8 to be non-Shakespearean, and we either agree or cannot tell except in the cases of 4.3, 4.4, and 4.6 (we say Shakespeare). They find 5.1 to be Shakespearean, and we cannot tell. They find 5.2 to be non-Shakespearean, and we agree. They find 5.3, 5.4, 5.5, 5.6, and 5.7 to be Shakespearean, and we either agree or cannot tell except in the case of 5.7 (we say Marlowe).

The main conclusion of our experiments with the *Henry VI* plays, then, is that recent claims for Marlowe's hand in them are corroborated. That independent teams of investigators find this to be the case using entirely different methods that are, as far as we can tell, impervious to mere literary impersonation is a strong reason to accept that these plays contain measurable amounts of Marlowe's writing. Just how this came about cannot be determined by our method, since Shakespeare taking over and rewriting a play first written by Marlowe (or Marlowe and others) would, by our method, test much the same way as a play that Shakespeare and Marlowe actively cowrote. The presence of Marlowe in these plays, however, is now undeniable.

VI. Other Recent Questions in Shakespearean Authorship

Our method also confirms the most widely agreed upon recent attributions in Shakespeare studies. Shakespeare collaborated with Middleton on *Timon of Athens* and with Fletcher on *Henry VIII* and *The Two Noble Kinsmen*. Middleton's alleged adaptation of *Macbeth* and *Measure for Measure* would, according to its proponents, have left in those plays samples of his writing that are too small to test by our method. Our results on other controversial recent claims may be usefully summarized here.

We agree with those who find that Peele wrote Act 1 of *Titus Andronicus* and that Shakespeare wrote Acts 2, 3, 4, and 5. Looking at each scene, we find

[16] Hugh Craig and John Burrows, "A Collaboration about a Collaboration," 59.

1.1 to be Peele's, 2.1 and 4.4 to be undecidable between Shakespeare and Peele, and the remainder of the play to be Shakespeare's. Independent studies, recently summed up by Jackson, have suggested that the central part of the anonymously published *Arden of Faversham*, including scene 8 (the quarrel scene), is by Shakespeare.[17] We agree, finding Shakespeare ahead of our seven other candidates for each act. The two main alternative candidates that have been offered for authorship of the play are Kyd and Marlowe, and because Kyd has just one reliably attributed play, *The Spanish Tragedy*, we possess too little material to test for him with our method. This would be true even if we tentatively accepted *Soliman and Perseda* as Kyd's play too.

Applying the techniques of Burrows and Craig, Timothy Irish Watt found that Shakespeare wrote the countess scenes of *Edward III*, meaning scenes 2 and 3 in the *Oxford Complete Works* (or 1.2, 2.1, and 2.2 in other editions).[18] Our act-wise analysis gives Act 1 to Marlowe and Acts 2 and 4 to Shakespeare. It has no clear answers for Act 3 (most like Shakespeare, Marlowe, and Peele in that order) and Act 5 (most like Shakespeare, Peele, and Greene in that order). Putting Shakespeare against Marlowe in a scene-wise analysis gives Marlowe 1.1, 1.2 (by the slightest of margins), 3.1, 4.1 , 4.7, and 4.8 and Shakespeare 2.1, 2.2, 3.2, 3.3, 3.4, 3.5, 4.2, 4.3, 4.4, 4.5, 4.6, 4.9, and 5.1. When Thomas Kyd's *Spanish Tragedy* was reprinted in 1602, it contained 320 lines, grouped into five Additions, that were not present in previous editions. Independent studies have recently claimed that Shakespeare wrote the Additions.[19] At the end of the first edition of Shakespeare's Sonnets appears a 329-line poem called *A Lover's Complaint*, the authorship of which has long been disputed with the most recent contribution (in favor of Shakespeare's authorship) coming from Jackson.[20] The Additions to *The Spanish Tragedy* and *A Lover's Complaint* are small samples of writing, and at this scale the accuracy of our eight-author attribution test falls to around 75 percent. That is, our test gives an incorrect attribution for every three correct attributions. Nonetheless, for the sake of completeness we mention that our test gives the Additions to *The Spanish Tragedy* to Shakespeare, ahead of Jonson by a small margin, and gives *A Lover's Complaint* to Chapman, ahead of Shakespeare by a small margin.

It must be remembered that for each text we have attributed our pool of candidates is limited to those whose works we know about and which survive in

[17] Jackson, *Determining the Shakespeare Canon*, 9–126.

[18] Timothy Irish Watt, "The Authorship of *The Raigne of Edward the Third*," in *Shakespeare, Computers, and the Mystery of Authorship*, 116–33 (see n. 4).

[19] Hugh Craig, "The 1602 Additions to *The Spanish Tragedy*," in *Shakespeare, Computers, and the Mystery of Authorship*, 162–80 (see n. 4); Brian Vickers, "Identifying Shakespeare's Additions to *The Spanish Tragedy* (1602)."

[20] Jackson, *Determining the Shakespeare Canon*, 129–218.

canons large enough for both random variation and systemic bias—most obviously with literature, the biases of genre—to cancel one another out. As we have observed, the canons of Kyd and Nashe are too small for our method, excluding them from our conclusions about authorship. More generally, it should be borne in mind that all of our conclusions are discriminations between the authors we are able to test.

That the technique presented here confirms a number of recent authorship-attribution findings made by entirely different methods should be of comfort to all investigators of these problems. In general, the more our independent studies converge on certain conclusions, the greater the likelihood that those conclusions are correct; when they use different methods, the errors of one investigator or team are not inherited by another. There is no reliable way to calculate exactly how much more likely it becomes that an assertion is true once multiple independent studies start to support it. But we can say how likely it is that certain outcomes could be reached by chance alone, and the odds for this study are comfortingly small.[21]

No one knows why methods that count frequencies of common words are able to distinguish authorship, and we offer no explanation for why our method of measuring their proximate adjacency is equally successful. We do not assume that either kind of choice is consciously made by authors, but it is intuitive that the relative placing of words—the ways in which "one word looks for another," in Vickers's phrase—would complement the study of overall word frequencies as we inch our way toward a better understanding of literary creativity.

VII. A Note on the Provenance and Processing of Electronic Texts

All of our testing is performed using electronic texts from the LION database supplied by the online digital publisher Chadwyck-Healey. For Shakespeare's plays, we prefer the versions that appeared in the 1623 First Folio over preceding quarto editions. We use the Folio's versions because they were produced in a

[21] There is about a one-in-a-hundred chance of our results aligning with existing claims as much as they do if our methods were no more reliable than random guesses of authorship. We may place some number on this chance by considering that for each play scene a random guess would give us a 50 percent chance of agreeing with the existing claim. For instance, in *3 Henry VI*, there are twenty-eight scenes, of which we agree with Craig on sixteen, disagree on six, and do not make any claims on the last six. In mathematical terms, agreeing at least sixteen times out of twenty-two, if we were just guessing randomly with a 50 percent chance of agreement, would occur with a probability of 0.85 percent, using what is called the binomial distribution in probability. We do not make direct comparisons for *2 Henry VI*, and in the case of *1 Henry VI* we find that there is about a 2.5 percent chance that by simply guessing we would align to the degree that we do. These probabilities are all well within the 5 percent threshold that is commonly used to mark statistical significance.

single printshop within a fairly short period (the years 1622–23) by a single team of compositors and proofreaders. For some of Shakespeare's plays, editors have preferred a preceding quarto edition because the Folio appears merely to reprint it, but there is sometimes a plausible (and as yet unresolved) argument for the Folio acquiring additional fresh authority in these cases.

The differences between good quarto texts and the Folio are in no cases large enough to substantially affect our results. The 1623 Folio texts are reasonably consistent in regard to spelling and layout, and the feature we detect—the proximities of function words—is relatively immune to depredations such as the expurgation of oaths and other kinds of censorship that are known to affect some plays in the collection. For other writers and for suspected Shakespeare works not present in the 1623 Folio, we choose the earliest edition offered by LION.

We perform minimal preprocessing of the LION electronic texts, relying instead upon on-the-fly discrimination of the features in which we are interested. We do not normalize or modernize the variant spellings of words that are now consistently spelled one way in English, and naturally this means that we treat words as if they were merely strings of characters. In English, the three letters *r-o-w* can stand for a verb for manually propelling a boat, a noun for an argument, a noun for a horizontal line of objects, and for other words besides. Unless texts are comprehensively tagged for lemmatization and morphosyntactic discrimination, computers can only "see" them as strings of letters and not as words in the sense that linguists mean. The variant spellings of early modern English present similar opportunities for misclassifying linguistic data. These limitations are not significant barriers to authorship discrimination for two reasons. The first is that for the most part the spellings of words found in early modern printed books are those chosen not by their authors but by their compositors who set the type, and these men were free to exercise their own preferences. The second is that these misclassifications affect all authors equally so that with large sample sizes, as used here, the distortions cancel one another out.

Our on-the-fly discrimination of textual features covers the following. We detect the beginnings of speeches by the white space that LION puts between them, and we do not count as falling within our moving window any adjacency that spans a speech break: this is the only segmentation we impose. We do remove all stage directions but not speech prefixes or act/scene division markers because they do not in any case contain the function words we are counting. Where there is a danger that an abbreviated speech prefix might be mistaken for a function word, as with the name "Anne" being abbreviated to "An," we check in advance and discount such cases.

Appendix 1: Function Words

(a) Full list of function words used in this method[22]

a	between	instead	part	towards
aboard	beyond	into	past	under
about	bit	it	pending	underneath
above	both	its	per	unless
absent	but	itself	pertaining	unlike
according	by	less	plenty	until
accordingly	can	like	plus	unto
across	certain	little	regarding	up
after	circa	loads	respecting	upon
against	close	lots	round	us
ahead	concerning	many	save	used
albeit	consequently	may	saving	various
all	considering	might	several	versus
along	could	minus	shall	via
alongside	couple	more	should	view
although	dare	most	similar	wanting
amid	despite	much	since	what
amidst	down	must	so	whatever
among	due	near	some	when
amongst	during	need	somebody	whenever
an	each	neither	something	where
and	either	nevertheless	spite	whereas
another	enough	next	such	wherever
any	every	no	than	whether
anybody	everybody	nobody	that	which
anyone	everyone	none	the	whichever
anything	everything	nor	them	while
around	except	nothing	themselves	whilst
as	excluding	notwithstanding	then	who
aside	failing	of	thence	whoever
astraddle	few	off	therefore	whom
astride	fewer	on	these	whomever
at	following	once	they	whose
away	for	one	this	will
bar	from	onto	tho	with
barring	given	opposite	those	within
because	heaps	or	though	without
before	hence	other	through	would
behind	however	ought	throughout	yet
below	if	our	thru	
beneath	in	out	till	
beside	including	outside	to	
besides	inside	over	toward	

[22] The particular set of function words used for each attribution is arrived at by removing from this list those words that are least useful in discriminating the styles of the candidate authors in each case. The words are presented here in modern spelling, but a limited set of alternative spellings active in the early modern period—derived from the *Oxford English Dictionary* (*OED*) lists of spellings—is permitted for each headword here.

(b) Function words (100 in total) used in the attribution of full-length plays, determined in the training process

a	could	much	past	to
about	dare	must	shall	until
after	down	need	should	unto
against	enough	neither	since	up
all	every	next	so	upon
an	for	no	some	us
and	from	none	such	what
another	given	nor	than	when
any	hence	nothing	that	where
as	if	of	the	which
at	in	off	them	while
away	into	on	then	who
bar	it	once	therefore	whom
because	like	one	these	whose
before	little	or	they	will
both	many	other	this	with
but	may	our	those	within
by	might	out	though	without
can	more	over	through	would
close	most	part	till	yet

(c) Function words (76 in total) used in the attribution of individual acts, determined in the training process

a	for	none	some	us
about	from	nor	such	what
against	if	nothing	that	when
all	in	of	the	where
an	into	off	them	which
and	it	on	then	who
any	like	once	these	whose
as	little	one	they	will
at	many	or	this	with
away	may	other	those	without
before	might	our	though	would
both	more	out	till	yet
but	most	shall	to	
by	much	should	unto	
can	must	since	up	
could	no	so	upon	

(d) Function words (55 in total) used in the attribution of individual scenes, determined in the training process

a	for	no	some	upon
all	from	nor	such	us
an	if	of	that	what
and	in	on	the	when
any	it	one	them	where

as	like	or	then	which
at	may	our	these	who
away	more	out	they	will
but	most	shall	this	with
by	much	should	to	would
can	must	so	up	yet

Appendix 2: Texts Used to Construct Author Profiles[23]

(a) Texts used to construct author profiles for Shakespeare, Marlowe, Fletcher, Jonson, Chapman, and Middleton

William Shakespeare

All's Well That Ends Well
Antony and Cleopatra
As You Like It
The Comedy of Errors
Coriolanus
Cymbeline
Hamlet
1 Henry IV
2 Henry IV
Henry V
Julius Caesar
King John
King Lear
Love's Labor's Lost
The Merchant of Venice
The Merry Wives of Windsor
A Midsummer Night's Dream
Much Ado About Nothing
Othello
Richard II
Richard III
Romeo and Juliet
The Taming of the Shrew
The Tempest
Troilus and Cressida
Twelfth Night
The Two Gentlemen of Verona
The Winter's Tale

Christopher Marlowe

Doctor Faustus
Edward II
The Jew of Malta
The Massacre at Paris
1 Tamburlaine the Great
2 Tamburlaine the Great

John Fletcher

Bonduca
The Chances
The Faithful Shepherdess
The Humorous Lieutenant
The Island Princess
The Loyal Subject
The Mad Lover
Monsieur Thomas
The Pilgrim
Rule a Wife and Have a Wife
Valentinian
A Wife for a Month
The Wild Goose Chase
The Woman's Prize
Women Pleased

[23] The spelling of play titles for Shakespeare derives from Barbara Mowat, Paul Werstine, Michael Poston, Rebecca Niles, eds., *Shakespeare's Plays, Sonnets, and Poems* (Washington: Folger Shakespeare Library, n.d.), www.folgerdigitaltexts.org (accessed 26 August 2016). The spelling of play titles for all other playwrights derives from Alfred Harbage, *Annals of English Drama, 975–1700*, rev. by S. Schoenbaum (London: Methuen: [1964]). Due to space constraints, the extended titles have been omitted.

Ben Jonson

The Alchemist
Bartholomew Fair
Catiline's Conspiracy
Cynthia's Revels
The Devil Is an Ass
Epicoene
Every Man in His Humour
Every Man out of His Humour
The Magnetic Lady
The New Inn
Poetaster
The Sad Shepherd
Sejanus His Fall
The Staple of News
A Tale of a Tub
Volpone

George Chapman

All Fools
The Blind Beggar of Alexandria
Bussy D'Ambois
Caesar and Pompey
The Conspiracy and Tragedy of Charles Duke of Byron
The Gentleman Usher
An Humorous Day's Mirth
May-Day
Monsieur D'Olive
The Revenge of Bussy D'Ambois
Sir Giles Goosecap
The Widow's Tears

Thomas Middleton

A Chaste Maid in Cheapside
A Game at Chess
Hengist, King of Kent
A Mad World, My Masters
Michaelmas Term
More Dissemblers Besides Women
No Wit, No Help Like a Woman's
The Phoenix
The Puritan
The Revenger's Tragedy
The Second Maiden's Tragedy
A Trick to Catch the Old One
Your Five Gallants
The Widow
The Witch
Women Beware Women

(b) Texts used to construct profiles for Greene and Peele

Robert Greene

Alphonsus, King of Aragon
Friar Bacon and Friar Bungay
Orlando Furioso
The Scottish History of James IV

George Peele

The Arraignment of Paris
The Battle of Alcazar
Edward I
The Love of King David and Fair Bethsabe
The Old Wives Tale