“A cannon's burst discharged against a ruinated wall”: A Critique of Quantitative Methods in Shakespearean Authorial Attribution

DAVID AUERBACH

Abstract: Authorship studies has, over the last two decades, absorbed a number of quantitative methods only made possible through the use of computers. The New Oxford Shakespeare Authorship Companion presents a number of studies that utilize such methods, including some based in machine learning or “deep learning” models.

This paper focuses on the specific application of three such methods in Jack Elliott and Brett Greatley-Hirsch’s “Arden of Faversham and the Print of Many.” It finds that their attribution of the authorship of Arden to William Shakespeare is suspect under all three such methods: Delta, Nearest Shrunken Centroid, and Random Forests. The underlying models do not sufficiently justify the attributions, the data provided are insufficiently specific, and the internals of the methods are too opaque to bear up to scrutiny. This article attempts to depict the internal flaws of the methods, with a particular focus on Nearest Shrunken Centroid.

These methodological flaws arguably arise in part from a lack of rigor, but also from an impoverished treatment of the available data, focusing exclusively on comparative word frequencies within and across authors. A number of potentially fruitful directions that authorship studies are suggested, that could increase the robustness and accuracy of quantitative methods, as well as warn of the potential limits of such methods.

Contributor Biography: David Auerbach is the author of *Bitwise: A Life in Code* (Pantheon), written under a New America fellowship. After graduating from Yale University, he did graduate work on James Joyce and Denis Diderot while working as a software engineer and architect at Google and Microsoft, where he specialized in optimization and heuristics. He currently authors the Aleph-Bits column for *Tablet* Magazine. His writing has appeared in *The Times Literary Supplement, MIT Technology Review, The Nation, Slate, The Daily Beast, n+1,* and *Bookforum*, among many other publications.

Introduction

Due to exponential increases in computing power and data storage over the last sixty years, quantitative analysis is now capable of feats of which we could not even conceive prior to the advent of computer technology. Researchers have pointed out how sheer quantity of data can allow for new successes with analyses that previously delivered poor results: speech recognition, spam filtering, and image classification are only three particular domains where refinement of existing methods has yielded monumental improvements primarily due to a massive increase in the amount of data being analyzed.[1]

It makes sense, then, for authorship attribution to take note of these developments. Efforts such as the LION database of English literature and the Perseus Digital Library of Greco-Roman works have established convenient and standardized references for computational analyses. More recently, the computer-assisted approaches of scholars such as MacDonald P. Jackson and Brian Vickers have been joined by a more statistically-focused approach that begins with John Burrows' Delta tests and extends to the machine learning methods utilized by some of the contributors to the *New Oxford Shakespeare Authorship Companion*.

At first glance, such work would appear to be a welcome development. Human fallibility and subjectivity are difficult to remove entirely from any scholarly effort, and so neutral, quantitative analysis would seem to provide a means to arbitrate the disagreements with cold statistics. On further investigation, however, there are flaws that are significantly endemic to these new approaches. While this paper will treat the lack of analytical rigor that makes many of these results unconvincing, a close study reveals doubts as to whether such approaches are suitable at all to attribution studies, and the flaws revealed demand modification to the methodologies that are implicitly and explicitly proposed by these new studies. I do not claim that we must stick to existing methods; on the contrary, there is certainly much to be gained from advances in data science and learning, but only when the proper groundwork has been prepared. Rather than a carefully sifted set of findings, the efforts, as summed up by the *Companion*, appear more to be a gold rush where precious material cannot be distinguished from pyrite, and the latter dominates.

One central problem is the *opacity* of the methods: not only are the statistical methods frequently employed without sufficient reference to confidence and error estimates, but there is reason to doubt whether these estimates would make the methods suitable for such attribution studies in the first place. This is in large part due to the second central problem, which is the poverty of the input data. By restricting such analyses to a handful of primitive signals such as word frequency and word succession, many of these researchers end up coating fundamentally simple (and untenable) findings in a statistical glaze, disguising the *explanation* for the results in precisely regimented charts and tables. A shift in focus from presentation of results to methodological justification is required. As described below, the explanations for the results are usually far simpler than the arguments employed by hand-crafted authorship attribution studies. Consequently, these methods run the risk of taking a step backward rather than forward in attribution studies. What is needed, instead, is a groundwork for the future employment of such methods which insists on the transparency, suitability, and surety of any generated results. A provisional such groundwork will be given at the end of this paper.

[1] See, for example, F. Pereira, P. Norvig and A. Halevy, "The Unreasonable Effectiveness of Data," *IEEE Intelligent Systems* 24 (2009): 8-12. https://doi.org/10.1109/MIS.2009.36 [last consulted 24 December 2018].

[1] Taylor, Gary, Jowett, John, Bourus, Terri, and Egan, Gabriel, "General Editors' Preface: A *Complete Works*," in *The New Oxford Shakespeare Authorship Companion* (Oxford, 2017), vi.

The *New Oxford* Approach

The *New Oxford Shakespeare Authorship Companion* contains, in the words of its editors, “an extensive sampling of new research, employing a range of new data and new methods in an effort to resolve problematic cases that have, hitherto, eluded resolution.”[2] It also makes bold leaps into the employment of computational techniques, particularly those associated with machine learning, in order to suggest new approaches to tackling the contentious and difficult-to-arbitrate authorship claims around plays such as *Arden of Faversham*, *All’s Well That Ends Well,* and *Double Falsehood*.

Among the claims made by contributors to the *Companion* utilizing these new methodologies are:

1. Shakespeare had a definite hand in *Arden of Faversham*.
2. Christopher Marlowe is most likely to have written “the non-Shakespeare portions of *3 Henry VI*”.
3. Thomas Middleton revised and added material to *All’s Well* 4.3.

These are revelatory results, either resolving vexed questions or raising previously unexplored possibilities. All are announced with a high degree of confidence, and moreover, these claims are obtained with a minimum of semantic analysis; i.e., the analyses rely primarily on lexical features of the texts in question rather than the semantic meanings of the words or phrases treated. By itself, this focus is not novel; word frequency, rare phrases, and bigrams/trigrams have all long been in the toolkit of attribution studies. The difference lies in that these computational methodologies tend to make diagnostic generalizations about the entirety of a text (or a text segment) rather than about hand-picked individual passages or phrases. The strength of such computational analysis lies in such large-scale number crunching; its complementary weakness is that humans cannot curate individual data by hand, as doing so comprehensively would be prohibitively slow and error-prone. If, for example, an analysis attempts to determine an “authorial fingerprint” by looking at the various frequencies of individual words in an author’s works, a researcher should not and generally cannot manually massage the data to remove certain words from consideration or artificially increase the weight of other words. This is as it should be: the analyses only gain consistency and validity if they are conducted with total uniformity.

Yet such computational analyses can possess a certain holistic opacity, in which surprising or puzzling results do not easily admit to interrogation or even explanation. This paper aims to examine these computational methodologies in particular with an eye toward opening up their (sometimes cryptic) results. Many machine learning methods operate as functional black boxes, in which the confidence in results does not originate from a verified or even visible model, but from the successful application of a trained machine learning model to a set of known data. If a model returns correct results in known cases, it should do so in cases where the answer is not known, all other things being equal. The goal of this paper is to judge whether all other things are indeed equal, and so assess the confidence which one can have in the new results.

Word Frequency Analyses

Jack Elliott and Brett Greatley-Hirsch (Elliott and Greatley-Hirsch from here on) offer no fewer than four quantitative methodologies for authorship attribution in "*Arden of Faversham* and the Print of Many," all of which utilize computational analysis to compare word frequencies between *Arden* and selections from the corpora of contemporary authors.[3] Claiming that all four show indications of Shakespeare's hand in well over the majority of *Arden*'s scenes, Elliott and Greatley-Hirsch conclude, "It is impossible to reconcile the results we have found with a belief that Shakespeare had no hand in *Arden of Faversham*." This conclusion is only true, however, if the results are both valid and well-grounded, and there are reasons for doubt in this regard.

Via email, Greatley-Hirsch told me that the numerical results for the tests summarized in the chapter were no longer available.[4] He was also unable to provide the parametric constants used in the analysis, such as the crucial threshold parameter used for shrinking in the nearest shrunken centroid method.[5] As a consequence, I have been unable to determine the algorithms used with a sufficient degree of specificity to allow me to replicate Elliott and Greatley-Hirsch's analyses, and my own reservations regarding the algorithms, detailed below, precluded me from making confident choices when faced with algorithmic ambiguities. Even had I done so, I would not have had numerical results to compare against had my own results disagreed, and little way to determine the cause of any disparities. Elsewhere in the *Companion*, Anna Pruitt writes, "[G]iving readers access to the raw data should be a non-negotiable part of the scholarly acceptance of studies of this nature; without access to the data the experiments cannot be replicated to validate the study." Her prescription is wise.[6]

What remains, then, is a set of four attribution methods:

1. Delta
2. Nearest Shrunken Centroid
3. Random Forests
4. Zeta

All four are applied to counts of subsets of words on a per-play basis, which are then grouped by author and then compared against the results for the anonymous author of *Arden*. The first three methods will be discussed in this paper. (Pervez Rizvi has written an extensive critique of the Zeta approach and so I pass over that method here.[7])

Each attribution method was applied to three different subsets of words considered across all plays. The three methods of word selection are: all words; the 500 most frequent words; function words. While the three methods of word selection produce somewhat differing results under all attribution methodologies, the merits and defects of each are not at the

[3] Elliott, Jack, and Greatley-Hirsch, Brett, "*Arden of Faversham* and the Print of Many," in *The New Oxford Shakespeare Authorship Companion* (Oxford, 2017), 139-81.

[4] Email correspondence, June 4 to July 28, 2018.

[5] Nearest shrunken centroid analysis is described in Tibshirani, Robert, et al., "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *PNAS* 99 (2002): 6567-72; https://doi.org/10.1073/pnas.082099299.

[6] Pruitt, Anna, "Refining the LION Collocation Test: A Comparative Study of Authorship Test Results for *Titus Andronicus* Scene 6 (= 4.1)," in *The New Oxford Shakespeare Authorship Companion* (Oxford, 2017), 104.

[7] Rizvi, Pervez, 'The interpretation of Zeta test results', *Digital Scholarship in the Humanities*, 23 August 2018: https://doi.org/10.1093/llc/fqy038 [last consulted on 24 December 2018].

heart of the concerns of this paper. Greatley-Hirsch was only able to provide the word frequencies of the 500 most frequent words, so my analysis only utilizes that set, though much of the analysis would apply to any selection of words.

Rather, the germane question is to what extent “big data” quantitative methodologies drawn from statistical and machine learning techniques can be considered valid when applied to *any* textual data consisting of individual word frequencies. Such analyses as Elliott and Greatley-Hirsch’s disregard collocations, word order, and word function (except inasmuch as function is reflected in the subset of words chosen). The two questions, then, are whether such large-scale analysis on primitive lexical data can produce a valid authorial “fingerprint,” and if it can, whether the attribution methodologies employed are sufficient to produce such a fingerprint.

Delta

Elliott and Greatley-Hirsch use John Burrows’s Delta test as their first attribution method. Delta, proposed by Burrows in 2002, can be considered a progenitor of the newer techniques, having been one of the first analyses to use computers to analyze large sets of words in a text rather than hand-picked (and often hand-counted) subsets.

Elliott and Greatley-Hirsch imply that Delta is less consistently reliable than their other methods, as they do not apply it in cases where they are analyzing all occurring words. In those cases, “we omit Delta because it is inaccurate when dealing with infrequently occurring words.”[8] Unfortunately, Elliott and Greatley-Hirsch have chosen to apply Delta on corpora that are so small that many words even within the top 500 words in terms of frequency occur only rarely. The 500th most frequently occurring word across all considered texts, for example, “less,” does not occur at all in *Arden*, while occurring four times in *Richard III*, six times in *The Spanish Tragedy*, and once in *Edward II*. Even those words around the middle of the list only occur with roughly twice the frequency as “less,” already far less than the top word, “and,” which has hundreds of uses in each text and thousands across the entire corpus under consideration, two orders of magnitude greater than the count for “less.”

Burrows himself has cast doubt on the validity of Delta. In discussing Delta’s matching of the non-Shakespeare segments of *3 Henry VI* to Kyd, Burrows and Craig write:

> Clearly there are affinities between *The Spanish Tragedy* and *Soliman and Perseda* and *3 Henry VI* across the board. This is a persistent finding with Delta tests, as discussed in this chapter. It is not supported by our other tests, and we discount it in our conclusions, explaining it as a general likeness in dramatic texture which does not survive more targeted authorial testing.[9]

If in this case Delta is not sufficiently “targeted” to produce accurate results, then there need to be criteria for determining in which situations Delta can and cannot be considered valid. If “general likeness in dramatic texture” is sufficient to produce a false attribution,

[8] Elliott and Greatley-Hirsch, 155.

[9] Burrows, John, and Craig, Hugh, “The Joker in the Pack? Marlowe, Kyd, and the Co-authorship of *Henry VI, Part 3*,” in *The New Oxford Shakespeare Authorship Companion* (Oxford, 2017), 202.

the use of such a subjective measure makes it impossible to determine objectively when Delta can be trusted.

Burrows appealed to similarly subjective measures in his first paper on Delta: "Many of the worst [Delta] results can be attributed to the fact that, in one way or another, a given poem or group of poems is uncharacteristic of its author."[10] Yet Burrows begs the question, for Delta purports to determine authorship *by virtue of* characteristic similarity. If there are some unspecified, subjective stylistic criteria of "uncharacteristicness" that can always defeat Delta, then Delta can never be considered reliable. When Burrows writes that "[o]nly a genuine authorial factor could yield results like those we have," he is incorrect. Rather, he has measured a factor that loosely correlates with authorship in unspecified ways, and therefore one of limited prescriptive value unless the nature of the relationship between Delta and authorial factor can be unambiguously detailed.

For the two reasons of the infrequency of most of the words under consideration in *Arden* attribution and the general lack of accuracy in existing Delta tests, Elliott and Greatley-Hirsch's Delta results do not bear scrutiny. Their two Delta tests only considered nine candidates (Shakespeare, Kyd, Marlowe, Wilson, Nashe, Lyly, Peele, Lodge, and Greene) and three candidates (Shakespeare, Kyd, and Marlowe) respectively. Burrows's tests show only an 85 percent success rate for Delta placing the correct author in the top *five* candidates.[11] Therefore, Elliott and Greatley-Hirsch's findings that Shakespeare was the top-ranked candidate in these Delta tests are immaterial.

Random Forests

The Random Forests test is unusual in being in principle non-replicable, because the test is non-deterministic.[12] Following machine learning procedures, a number of "decision trees" are constructed that classify a text segment into one of a number of categories (in this case, authors) by asking various questions of it—in this case, again, questions exclusively relating to individual word frequency. Elliott and Greatley-Hirsch write:

> The tree consists of a set of connected rules that describe the data, such as 'when the frequency of occurrence is less that 0.5 for the word soldiers and is greater than 0.008 for the word blazon or is less than 0.03 for the word master then the text is by author A, otherwise not'. The algorithm generates these rules by repeatedly decomposing the training set data—that is, dividing and subdividing its records—until it finds features that correlate with authorship.

The significance of the name lies in the fact that multiple decision trees are constructed by arbitrarily withholding parts of the dataset in order to use such withheld segments (termed "hold-out" or "out-of-bag" segments) to test the accuracy of the trees constructed with the remaining data. Those tests determine the generalization error of the overall Random

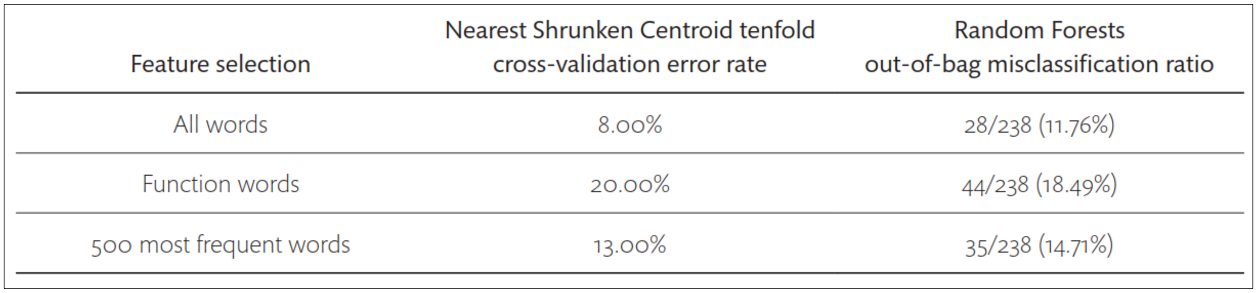
[10] Burrows, John, "'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship," *Literary and Linguistic Computing*, 17 (2002): 267–87.

[11] Burrows, ibid., 276-7.

[12] Greatley-Hirsch was unable to give me the intermediate or final results, such as variable importance, from the Random Forest tests he conducted. In the absence of knowing such parameters as the exclusion rate of the bootstrap data or the number of trees, any attempted replication would be invalid.

Forests test. Elliott and Greatley-Hirsch then apply the trees collectively to segments of *Arden* text, and each tree contributes one "vote" to the ultimate decision of authorship.

The success of Random Forests naturally depends in large part on how closely the opaque factor for which Random Forests tests correlates to the actual factor being sought. In his paper on Random Forests, Leo Breiman primarily concerns himself with demonstrating that Random Forests gives comparably accurate and more robust results than the deterministic Adaboost mechanism for evolving decision trees.[13] The absolute error rates of his tests vary wildly, naturally, depending on the size and nature of the data. While Random Forests may be a better choice for constructing authorship decision trees than decision tree algorithms that boost or otherwise rewrite the test data in constructing trees, the out-of-bag error rates Elliott and Greatley-Hirsch give are only middling, hovering between the 10 and 20 percent misclassification range in all of their tests.[14]



Feature selection	Nearest Shrunken Centroid tenfold cross-validation error rate	Random Forests out-of-bag misclassification ratio
All words	8.00%	28/238 (11.76%)
Function words	20.00%	44/238 (18.49%)
500 most frequent words	13.00%	35/238 (14.71%)

Table 1: Elliott and Greatley-Hirsch's "summary of error rates for Nearest Shrunken Centroid and Random Forests tests conducted on Arden of Faversham and the corpus of well-attributed, sole-authored plays, 1580–1594, using all words, function words, and 500 most frequent words."[15]

Yet the Random Forests error rate raises a puzzling anomaly in the results. More than Delta or Nearest Shrunken Centroid, Random Forests repeatedly decides for Shakespeare as the author of the entirety of *Arden*, which no scholar has maintained. When Elliott and Greatley-Hirsch applied Random Forests on all words and top-500 words in nine-author and three-author corpora (four tests total), all 35 segments of *Arden* were assigned to Shakespeare in each test except in one single case: when segment 17 was tested against the nine-author Random Forest trees using 500 most frequent words.

Breiman states that with regard to the error rate for a bagging classifier such as Random Forests, "the out-of-bag estimate is as accurate as using a test set of the same size as the training set."[16] Thus we can expect that the tests on *Arden* would have error rates between 10 and 20 percent, and so we would expect somewhere between 3 and 7 segments of *Arden* to be misclassified on average by each test, or between 14 and 28 misclassifications overall. And yet with one sole possible exception (segment 17 above), any and all such errors were made in Shakespeare's favour, a result which is statistically unlikely to have arisen by chance alone.[17] Even in the event that *Arden* had been entirely

[13] Breiman, Leo, 'Random Forests.' *Machine Learning* 45 (2001): 5–32.

[14] Elliott and Greatley-Hirsch, 159.

[15] Elliott and Greatley-Hirsch, 159.

[16] Breiman, 11.

[17] Elliott and Greatley-Hirsch's Random Forest tests on function words attribute five and four segments of *Arden* to Kyd, but this merely underscores how striking it is that Random Forests would produce uniform results in favour of one author.

composed by Shakespeare, one would naively expect that some segments still be (mis)classified as belonging to other authors.

It is difficult to speculate on the reasons for the Random Forests classifiers being slanted toward Shakespeare without a breakdown of the misclassifications made during classifier construction, but the apparent Shakespearean bias of the Random Forest classifiers casts doubt on those Shakespearean attributions which it does make, especially considering that Elliott and Greatley-Hirsch's other methods do not produce such uniformly Shakespearean results.

The greater question—to what extent individual word choice frequency is conclusively indicative of authorship—is one that will be treated subsequently in the discussion of Nearest Shrunken Centroid, which utilizes a less opaque method.

Nearest Shrunken Centroid

While Nearest Shrunken Centroid (NSC forthwith) utilizes the same raw data as the previous methods—single word frequency counts—it uses geometric distance in order to gauge the "distance" between works.

The method begins by conceiving of an n-dimensional space in which each axis represents the number of occurrences for a particular word. Thus, there are as many dimensions as there are words under consideration. Any text can be placed into this space at a single point, representing the number of occurrences for all *n* words under considerations. For example, in a 6-dimensional space where the dimensions respectively represent "The," "a," "dog," "cat," "around," and "ran," the sentence "The dog ran around the cat" would be represented by the point [2, 0, 1, 1, 1, 1].

Within this space, the works for each candidate author are divided into equal-length segments, which are then considered first separately and then together in order to produce a "centroid," which is a single point representing the average occurrence of each word for that author. Unknown texts can then be placed into the space, and the closest centroid point is deemed the winner and authorial candidate.

NSC makes some alterations in the calculation of the per-author centroid in order to take into account both the predominance of words and the variation (or standard deviation) of a given word across an author's works. Elliott and Greatley-Hirsch state their NSC methodology is taken from a statistical biochemistry paper, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," so what follows is based on the description of the method given there.[18]

A set of works by a particular author (or class, in genetic terms) is divided into *n* segments (or samples) of equal size, each containing the frequencies of *p* words (or genes). First, the *overall* centroid for all authors is calculated, giving a point representing the average occurrences for each of *p* words across all authors. Then, per-author centroids are calculated as *differences* in average occurrence per word. That is, these centroids do not represent absolute word occurrences, but the differences from the average occurrence of a

[18] Tibshirani, Robert, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu, 'Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression,' *Proceedings of the National Academy of Sciences of the United States of America* 99 (2002): 6567–72. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC124443/ [last consulted on 24 December 2018]. Elliott and Greatley-Hirsch do not describe the shrinkage method in detail, but its workings are crucial to understand how NSC produces the results it does, and so the shrinkage procedures described here are drawn from Tibshirani et al.

given word across all authors—whether a particular author tends to use a word more or less frequently than the average. These differences are then divided by the per-author standard error for that word in order to standardize the data.

In other words, NSC attempts to isolate particular words that a particular author *consistently* uses either more or less frequently than the mean usage across all authors. If the author's usage of a particular word is too inconsistent across segments, NSC will weigh that word less in its determination of its stylistic "fingerprint" for the author, which is a consequence of standardization. NSC goes further, however, in that a word may be eliminated from consideration altogether if (a) an author's average usage of a word is not sufficiently different from the average usage across all authors; or (b) the author's usage of the word is too irregular. Tibshirani et al. used this analysis to identify individual genes that, within a particular class, consistently deviated by a significant amount from the expression of that gene across all classes.

One consequence of this adjustment is that *any* word with inconsistent usage will contribute little role in NSC analysis. In practice, any word relevant to context, such as "murder" or "love" or "pleasure," will not play a significant part in NSC calculation unless an author used such a word with incessant regularity in each 2000-word (in Elliott and Greatley-Hirsch's case) segment of each of his plays under consideration. In practice, the restriction to consistent usage would remove many of the top 500 most common words (which include "slave," "money," and "mad") from impacting the analysis.[19]

Tibshirani et al.'s goal was to eliminate the vast majority of genes from consideration and isolate a small percentage of genuinely indicative genes. In the example they give, their NSC procedure takes a set of 2308 genes and yields only 43 genes (approximately 2 percent) whose per-class variations from the overall mean are sufficiently consistent and large to utilize in NSC calculations. In other words, determinations of similarity are ultimately based on a very small subset of genes. And though not explicitly stated by Elliott and Greatley-Hirsch, the resulting effect of the centroid shrinkage is to make authorial attributions based only on a small set of words used with consistent regularity by an author.[20]

Without knowing the threshold parameters Elliott and Greatley-Hirsch used, any attempts to identify decisive words are speculative, but some general conclusions can be drawn.[21] NSC ultimately determines a list of decisive words on a per-author basis. I conducted a NSC analysis on the subset of finalists for *Arden,* Shakespeare, Kyd, and Marlowe. NSC finds that some words, like "I" and "you," are relevant to all three authors, with Shakespeare consistently using both more frequently than Kyd or Marlowe.[22] The

[19] Elliott and Greatley-Hirsch write, "The refinement that gives Nearest Shrunken Centroid its name involves diminishing the significance of the counts for function words that are inconsistently used by authors," (Elliott and Greatley-Hirsch, 148), but this adjustment is performed on *all* words, not merely function words. In practice, NSC's elimination in favor of most consistently used words turns out to rely primarily on function words in any event, with a handful of exceptions: "sir" (favored by Shakespeare, avoided by Marlowe), "good," (favored by Shakespeare) and "heavens" (favoured by Kyd) were among the few non-function words to survive high thresholds in my own NSC tests.

[20] Again, the exact reduction depends on the now-unavailable NSC threshold parameter. But since the stated purpose of NSC is to exclude a high number of genes/words from consideration to isolate a handful of anomalies, there would be no point in using NSC over unshrunken centroids (which, as Elliott and Greatley-Hirsch say, are too sensitive to noise) if the reduction were not significant.

[21] Greatley-Hirsch reported that the internal parameters and numerical results that he had prepared for the NSC analyses were no longer available.

[22] My findings here agree with Elliott and Greatley-Hirsch, 170.

word “their,” however, is only relevant for Shakespeare: *Arden*’s underuse of “their” helps NSC classify it as Shakespearean, but a surfeit of “their”s would *not* contribute to adjudicating between Kyd and Marlowe, even though both also use “their” with higher than average regularity. It would only count against Shakespeare.

Some of this is no doubt contextual. Shakespeare’s consistently high usage of function word “she” no doubt owes to three of the four plays under consideration being *The Comedy of Errors*, *The Taming of the Shrew,* and *The Two Gentlemen of Verona*. Contrariwise, Marlowe’s consistently low usage of it in *Edward II, The Jew of Malta, The Massacre at Paris,* and *Tamburlaine* is also not surprising—like the fourth Shakespeare work *Richard III*, women are far less central to these Marlowe plays. But NSC takes this word to be decisive, and thus *Arden*’s slightly below-average usage of “she” plays a part in NSC’s attribution of two *Arden* segments to Marlowe. Ironically, these segments contain scene 8, which has long been held (by Jackson and others) to be the *most* likely Shakespearean scene of *Arden*, despite its paucity of “she”s. A burst of six “she”s from Franklin in scene 9, however, tilts NSC back toward Shakespeare in classifying subsequent *Arden* segments.

The same holds true for the non-function word “mistress,” used more often by Shakespeare in the three comedies than by the other two playwrights anywhere. *Arden* also contains a surfeit of uses of “mistress,” but again, not in scene 8. Similarly, the three Shakespearean comedies use “sir” with such high regularity that the word’s near-absence in *Richard III* does not obviate its significance to the NSC classifier. The near-absence of “sir” and “mistress” from scene 8 of *Arden* is a significant factor in why NSC classifies it as Marlowe’s and Kyd’s in several of Elliott and Greatley-Hirsch’s analyses, while still classifying much of the rest of *Arden,* in which “sir” occurs somewhat more frequently, as Shakespeare.

Likewise, the three Kyd plays under consideration use “heavens” frequently and consistently enough that the lower than average use of “heavens” in *Arden* is a significant factor in Kyd’s rejection by NSC. By itself this is an interesting finding, but whether it merits the importance assigned to it by NSC is not established, nor is its importance made clear by Elliott and Greatley-Hirsch’s results. The same applies for the close succession of four instances of “sir” in scene 14 that help to clinch NSC’s conclusion that Shakespeare wrote the segments containing that scene, despite the lack of “she” in scene 8, which is included in some of those same segments. In other words, an influx of “sir” serves as a counterweight in the calculations to tilt the balance back to Shakespeare despite NSC taking a paucity of “she” as an indicator of Kyd’s hand. These are only aspects of the overall calculation, but nonetheless, the overall attribution remains a matter of tallying each of these per-word factors and seeing which candidate ultimately comes out ahead. Some words count for more than other words, but sufficiently aberrant frequencies for one word can counter aberrant frequencies in the opposite direction for other words.

In sum, though NSC may give statistically semi-reliable attributions, the factors it uses in its determinations are highly questionable and likely to be biased. As in the case of “she,” both the training data and the segments to be classified are simply not large enough to avoid the problem of using suspect indicators in authorial determinations.

Below is a summary of the words used in NSC analysis given a moderate threshold value (2.25). Lower threshold values caused NSC to take into account words with sufficiently low occurrences (nearing one or zero occurrences per sample on average) that conclusions of consistent over- or underrepresentation seemed impossible to justify. Only the top 100 most frequently occurring words occur two or more times per sample on average. Those words that survived the higher threshold of 3 have been boldfaced.

	consistent overrepresentation	consistent underrepresentation
Shakespeare	**I**, a, **you**, me, her, she, **sir**, good, am, say	**and**, **the**, **of**, shall, our, their
Kyd	that, but, death, whom, fortune, **heavens**	**you**, **your**, have
Marlowe	**and**, **the**, **of**, **shall**, we, let	her

Table 2: Words which under Nearest Shrunken Centroid analysis showed a consistent over- or under-representation within the works of a given author, relative to the mean use of that word across all authors. Nearest Shrunken Centroid uses the t statistic as a threshold for determining whether a particular word should be factored into its analysis. Boldfaced words surpassed a t statistic threshold of 3, while non-boldfaced words surpassed a t statistic threshold of 2.25.

This shortlist gives an indication of which words played the strongest roles in classifying *Arden*. In particular, while "you" is overrepresented in *Arden* overall relative to the average of works considered, it is underrepresented in scene 8, a major factor in NSC tipping toward Kyd, who consistently uses "you" significantly less than the average for the three authors. The same holds for "and," the single most frequent word across the works under consideration and one underrepresented in Shakespeare and overrepresented in Marlowe (Kyd's usage is close to the mean). While *Arden*'s overall frequency of "and" closely matches that of Shakespeare (approximately 62 usages every 2000 words), the first 259 lines of scene 1 contain "and" at Marlowe's higher average rate (approximately 85 usages every 2000 words). Not coincidentally, NSC repeatedly finds that Marlowe wrote this segment

Similarly, NSC at times assigns scene 8 to Marlowe and Kyd, in large part because its higher-than-average frequencies of "and," "the," and "you" move *Arden*'s position away from Shakespeare's shrunken centroid. Whether these frequencies are enough to counter MacDonald P. Jackson's analysis of Shakespearean parallels in scene 8, and his opposing contention that scene 8 is *more* Shakespearean than other parts of the play, is a question which Elliott and Greatley-Hirsch do not address.

The variation of "and" appears to exercise substantial influence in the attribution of the 2000-word *Arden* segments, pulling the NSC classifier toward and away from Shakespeare depending on the segment under consideration. Here are the mean number of occurrences of "and" per 2000 words across the known plays under consideration per author:

author	mean occurrences of "and" per 2000 words
Shakespeare	61.3
Kyd	69.7
Marlowe	84.2

Table 3: Mean usage of "and" within a 2000-word segment of a given author.

Now consider the division of *Arden* into 2000-word segments and the counts Elliott and Greatley-Hirsch give for "and" in each:

segment	*Arden* line-number range	occurrences of "and"
1	2-259 (Scene 1)	85
2	259-520 (Scene 1)	59
3	520-794 (Scenes 1–3)	57
4	794-1067 (Scenes 3–4)	53
5	1069-1334 (Scenes 4–8)	77
6	1334-1599 (Scenes 8–9)	59
7	1600-1876 (Scenes 9–13)	45
8	1877-2143 (Scenes 13–14)	77
9	2143-2423 (Scene 14)	52

Table 4: Occurrences of "and" per 2000-word segment of Arden*. (Source: internal data provided by Greatley-Hirsch.)*

It is precisely in those textual ranges where the occurrences of "and" jumps into Marlowe's range—segments 1, 5, and 8 above—where NSC makes its non-Shakespearean attributions. In most NSC tests of these segments or of segments significantly overlapping with them, there were not enough Shakespearean counts of other significant words for Shakespeare to "win" the attribution.

One notable exception is the NSC function word test of segment 8, which does find for Shakespeare. Of the 26 words that show consistent deviation beyond the 2.25 threshold value, seven do not belong to the list of function words Elliott and Greatley-Hirsch specify: sir, good, say, death, fortune, heavens, and let. It is therefore likely that the absence of these words is largely responsible for the shift in NSC attribution of *Arden* segments 28 and 29 (which both overlap scenes 13 and 14 in part, lines 1877-2212) from Kyd to Shakespeare when only function words are considered. The data support this contention. The words "sir," "good," and "say," all used with above average regularity by Shakespeare, appear at average or below-average levels in segment 28 (segment 8). When these words are excluded from consideration, their "evidence" against Shakespeare's hand disappears and help Shakespeare to prevail.

Yet *Richard III,* unlike the three Shakespearean comedies, does not contain a surfeit of instances of "sir". As with "she," the consistent overrepresentation of "sir" NSC finds in Shakespeare is a consequence of the homogeneity of the three comedies, not of Shakespeare *in toto*.

The greater question is whether the frequencies of words like "and," "sir," and "you" should play as strong a role in attribution as they do in the NSC analysis. These choices emerged by fitting NSC to the data to produce the lowest error rate—but the fact that the error rate was so much worse than for its original genetic application raises the possibility that NSC is the wrong tool. But this judgment is better made with an understanding of the criteria NSC determined to be most relevant for its attributions in this particular case—data that Elliott and Greatley-Hirsch do not provide and that had to be reverse-engineered. I argue that more than simply producing results, such large-scale quantitative analyses must also explicitly produce their criteria by which a classification was made, and this information should accompany any publication of the results.

Words are not genes, and we know far more about how words relate to each other than how genes do. The use of a gene fitting NSC algorithm to data about which far more context is available—only to yield worse results than the algorithm does on genes—seems a step backwards from the rigorous stylometric work that has already been done.

The Limits of Word Frequency

The limitations of Elliott and Greatley-Hirsch's word-frequency tests are evinced by their own results: the high error rates and opacity of the attribution algorithms render the results unconvincing. The argument that the tests are mutually reinforcing fails as well, as the tests are not independent of one another but in fact are closely related, all tests using the same fundamental word frequency data.

None of these limitations are present in the original NSC method and results as presented by Tibshirani et al., which raises the question of why the standards of that paper were not applied in Elliott and Greatley-Hirsch. In their essay there is a striking contrast between the apparent statistical sophistication of the methods and the ultimately primitive nature of the emergent criteria. Across all analyses, the results end up depending on a small number of word frequency differences. These criteria are not unrelated to the authorial factor, and in fact would have been interesting results had they been presented in the chapter, but they are far from decisive. The only justification given for these methods is the error rates, which in themselves are not reassuring, always running upwards of ten percent *on their own training data.* The tests dress up word frequency counts in opaque computational methods, but do not bring much new to the table. The most genuine innovation is that which lies within statistics, the nature of self-validation, but statistical correlation, particularly loose statistical correlation, in no way constitutes a theoretical validation of the method utilized.

The fundamental problem seems to me to be opacity rather than methodology per se. While Elliott and Greatley-Hirsch give an overview of their methods, there is no transparency as to why any method produced the results that it did. In the case of Delta and Random Forests, the methods remain opaque. The Zeta results, as Rizvi has shown, were incorrectly interpreted. When a method is investigated, as was done here with NSC, the ultimate decisions turn out to have been quite primitive, but the criteria for making them was constructed mechanically rather than by hand.

This examination of the internals of the NSC's analysis helps to depict the limits of analyses based on word frequency. All four of the methods Elliott and Greatley-Hirsch employ depend on comparing the word frequencies of 2000-word segments of *Arden* to those of the candidate authors, utilizing a variety of statistical tools. The implicit contention, taken for granted by Elliott and Greatley-Hirsch, is that single word frequencies are sufficient to establish a high degree of confidence in authorial attribution. Had their error rates been comparable to those of Tibshirani et al. in matching genomes, there might have been support for this contention. But given the 10 percent and higher rates of error in training classification, Elliott and Greatley-Hirsch's chapter rather shows there is good reason *not* to consider word frequency as anything beyond loosely indicative. Perhaps a statistical method exists based on word frequencies that produces better results; if so, Elliott and Greatley-Hirsch did not find it. I would argue, however, that given the disproportionate emphasis placed on word frequency approaches and the lack of high-confidence results in all the methodologies considered here, research efforts would be better spent in engaging with data beyond word frequencies.

Rebuilding the Foundation

As the example of Elliott and Greatley-Hirsch shows, computational authorship methodologies, which apply large-scale quantitative methods to problems of authorial attribution, face several recurring pitfalls:

1. The disregard of syntactic, semantic, and even lexical characteristics of a text, in favor of pure numerical measurements of base frequency.
2. Lack of differentiation between more and less meaningful authorial markers, weighing all markers by some uniform baseline metric (e.g., frequency).
3. The failure to establish or justify an underlying model dictating the validity of the analysis.
4. The conflation and confusion of discrete technical terms.
5. Failure to publish actual results beyond non-quantitative summaries, in particular confidence estimates for results.
6. Selective and subjective utilization of results to strengthen a particular case, disregarding contravening results.
7. Exaggerated claims for the success of a given method and failure to compare with other real or potential methods.

Taken together, these concerns serve to inject a significant element of doubt into the results of these computational studies as well as their role in the conclusions of the *Companion* more generally. I stress that I am agnostic concerning all of the conclusions proposed by these studies, other than the conviction that the stated level of confidence in those conclusions is unwarrantedly exaggerated overall. Given the tangled and zigzagging history of authorial attribution, in which certain conclusions are often undermined only to be resuscitated later in light of new evidence or methodologies, a greater degree of tentative humility would seem to be required.

It is an unfortunate tendency of the computer age that quantitative results frequently receive an automatic level of credence simply by virtue of their supposed neutrality and objectivity. This process, which I term "data laundering," causes the assumptions behind the modelling and selection of the data to be downplayed or ignored altogether—indeed, the absence of explicit models remains one of the greatest pitfalls of machine learning. Yet the lack of justifications for quantitative models can extend even to the fundamentals of the analysis, such as word frequency or syntactic redundancy. Richard Biernacki has criticized this sort of false objectivity:

> It is more transparent, therefore more faithful to inquiry, to assume radical difference in a population than to rush toward aggregating modern "facts" out of corpora whose members are artificially assumed to have homologous structures.[23]

[23] Biernacki, Richard, *Reinventing Evidence in Social Inquiry: Decoding Facts and Variables* (Palgrave Macmillan, 2012), 5.

In many of the authorship tests above, the “members” of which Biernacki speaks are the individual words themselves, treated as indistinguishable units with minimal semantic content. The quantitative methods discussed here produce valid results only on the assumption that all other things not treated are considered equal, and Biernacki’s point, as well as mine, is that all other things are not equal.

In light of the dubious nature of the results, we are struck by the impoverishment of restricting so much of the quantitative research to lexical units. Indeed, we wonder whether quantitative lexical analyses can ever gain the level of certainty required. The error rates of Elliott and Greatley-Hirsch’s fitting tests suggest that in the case of word frequencies, the data may simply not allow a sufficient degree of confidence to produce conclusive results. The decision to treat word frequencies in isolation is, from this standpoint, a perverse one. Given the many lexical and syntactic features available in the texts under consideration, the myopic focus on word frequency is counterproductive, particularly when the above results show little sign that such a focus fails to produce convincing results.

The following are only a handful of the purely syntactic and lexical features that could be incorporated into a single analysis in search of an author’s “blueprint”:

- Sentence lengths
- Location within the text
- Grammatical construction order and formulation
- Word adjacencies/clusters
- Interjections
- Punctuation
- Spelling preferences
- Bigrams, trigrams, et al.
- Character speech ratios
- Scene length and character representation in scenes

While any single feature in isolation may not be capable of providing high-confidence attributions, they may be capable of providing more persuasive results when combined. There may be more definite statistical *relationships* between some of these features that only emerge with the endless comparison and analysis that computers can provide. The strength of machine learning algorithms is the ability to generate and test countless models in competition with each other. A more sophisticated usage of machine learning would allow for different features to be included and excluded in due course in search of a set of criteria which provide the greatest distinguishing features for an author. (Nevertheless, the size of some corpora, such as Kyd’s, may remain too small to generate sufficiently high confidence intervals.)

Even if the resulting attributions are indecisive, the constructed criteria may prove useful for other analyses. While Kyd’s seeming affection for the words “heavens,” “fortune,” and “death” is not sufficient to attribute *Arden*, such a result is nonetheless relevant to Kyd scholars, making it all the more unfortunate that Elliott and Greatley-Hirsch did not note it. If rigorous application of supervised learning networks does not produce sufficient certainty in authorship attributions—a real possibility given the comparative paucity of data available—the most beneficial use of such tools may instead be in isolating unusual and disproportionate features of an author or text that scholars may not have noticed.

One can also conceive of extending such analyses to semantic features, provided those semantic features can be explicitly provided as annotations to the texts. Computers are not capable of recognizing image clusters or rhetorical figures with any degree of reliability, but they are capable of analyzing their presences if such information is provided as metadata. If LION texts were to be consistently and thoroughly annotated with markers for parts of speech, word etymologies, metaphors and metaphrands, imagery, other rhetorical figures, computational analyses could take such metadata into account conjointly with syntactic in lexical markers in searching for a best-fit attribution algorithm. There is a subjective component to all such labeling, of course, but recognizing this limitation is only akin to recognizing the subjective component to *all* such analyses. In the analyses considered above, the subjective component was present, for example, in the decision to assign such an unwarranted degree of importance to word frequencies to begin with.

Particularly with such a small corpus size and with such a low tolerance for error, statistical learning methods must remain as transparent as possible, requiring that researchers understand the decisive features causing their methods to produce the results that they have. The internal rationales for such results must also be presented in all openness and honesty, which entails making available in perpetuity both the exact input datasets and replicas of the resulting learning networks. Because of the unstable and dynamic nature of such networks, the only way to adjudicate disagreements between replications of studies will be if the original experiment can be precisely rerun. Fortunately, given the storage capabilities available today, the cost of storing and maintaining such an archive would be trivial, so long as it is competently administrated. Only then can such results reveal meaningful patterns within centuries-old texts, rather than serve as vague and ineffectual “cannon blasts”—as *Arden* has it—at the surfaces of opaque monuments.