The Writeprints of Man:
a Stylometric Study of Lafayette's Hand in Paine's 'Rights of Man'.

(Richard Forsyth, independent researcher; David Holmes, George Mason University, USA.)

[Please cite as:
Forsyth, R.S. & Holmes, D.I. (2018). The writeprints of man: a stylometric study of Lafayette's hand in Paine's 'Rights of Man'. *Digital Humanities Quarterly*, 12(1).
]

Abstract
Thomas Paine's *Rights of Man* (Part I, 1791; Part II, 1792) is one of the most influential political tracts ever written. Recently, Clark (2015) has queried the long-established notion that Part I is the work of a single author, Paine himself. Clark argues that a passage of approximately 6000 words just prior to the "Declaration of the Rights of Man and Citizens", near midway through the book, was written by Gilbert du Motier, Marquis de Lafayette, and inserted by Paine into his book, perhaps with some editing. Clark's assertion rests mainly on judgement of the tone and content of the queried passage, as well as a hint in a letter by Lafayette to George Washington. This paper presents the results of a stylometric study of this question. Three computational approaches to stylistic analysis -- one involving a consensus of six attribution techniques -- were applied to the text of the *Rights of Man*, Part I, along with a corpus of comparison texts from the same period. Our findings tend to corroborate Clark's contention, as well as identifying two other short passages in which Lafayette may have had a hand, thus rendering this important document ripe for re-interpretation.

Keywords: Authorship Attribution, Co-Authorship, Computational Stylistics, Revolutionary Writings, Stylometry, Text Classification.

1. Introduction

More than the writings of any other writer, those of Thomas Paine (1737-1809) illustrate the transformation in the meaning of the term "revolution" from the mid-eighteenth century to the beginning of the nineteenth century; indeed Paine could lay claim to be the world's first truly international revolutionary. Christopher Hitchens (Hitchens, 2006), in his portrait of Paine (*Thomas Paine's Rights of Man*, 2006) argues for three possible sources for the fermentation at work within him: (i) his upbringing with a Quaker father in the quasi-feudal Norfolk town of Thetford; (ii) the time that Paine spent on the lower deck of the ship *King of Prussia* in 1756 surrounded by a Royal Navy crew of nonconformists; (iii) the influence of the London scene of taverns and the working-man's lecture hall. On presenting himself to Benjamin Franklin and receiving a rather tepid letter of recommendation, Paine set sail for Philadelphia in 1774, arriving just as the colonial crisis with the British motherland was mounting. From his own experience of being an ill-used civil servant as an Excise officer, Paine was readier than most to advocate separation and independence.

1.1 Paine's Career as a Writer
With the help of Dr. Benjamin Rush, who had suggested the title and also found him a printer, in 1776 Paine published *Common Sense*. The result was a best seller on a scale hitherto unknown. Linking an earthy appeal with an inspirational style, it did more than any other publication to persuade America of the justice and necessity of independence. Thereafter, Paine became a major figure in the pamphlet and newspaper controversies of the Revolution, bringing his pen to bear

whenever he felt the American cause needed upholding, not least with his rallying cry in *The American Crisis*.

In the five years after the end of the American Revolutionary War, Paine more or less left politics behind and became absorbed in a series of scientific experiments, in particular the design and construction of a single-arched iron bridge. In 1787, again following the advice of Benjamin Franklin, who told him he would do well to seek sponsors for his bridge in either Paris or London, Paine returned to Europe at the very time when revolutionary and radical pressures were building there. He did not lack for well-placed friends. His admirer Thomas Jefferson had been appointed to be American Minister to France, and his old friend The Marquis de Lafayette, wreathed with American laurels, was also at his disposal. Lafayette kept a copy of the American Declaration on one panel of his study, leaving the opposite panel empty hoping it would be adorned one day by a similar French one (Hitchens, 2006, p. 47). Lafayette had been prominent in the parliamentary exchanges that led to the isolation of the French ruling dynasty, was also the commander of the National Guard and in the thick of the street demonstrations that had culminated in the fall of the Bastille on 14th July 1789. The French Revolution, which split British politics in several directions, struck Paine as a natural extension of the American one.

Paine's *Rights of Man*, written as a defence of the French Revolution, was issued in two parts. The first part (March 1791) was dedicated to George Washington and the second part (February 1792), calling for the spreading of the French Revolution across mainland Europe, to Lafayette. At this point, the gloves started to come off with Paine receiving a summons to appear in a British court and answer a charge of seditious libel. The hand-picked jury outlawed him in his absence and Paine never again returned to England. His time in France also turned out to be a hazardous enterprise when he spoke out against the execution of Louis XVI and backed the Gironde faction rather than the Jacobins. He was arrested in December 1793, held in the Luxembourg prison, and only narrowly escaped the guillotine due to a warder's error. The arrival in Paris of a new ambassador -- the future president James Monroe -- saw Paine's release in November 1794. His remaining years in France give the impression of the sour aftermath of a love affair (Hitchens, 2006, p. 63). In September 1797 the irruption of the army into politics was confirmed by a military coup and in November 1799 Napoleon arrogated all power to himself, proclaimed himself "First Consul" and announced that the Revolution was at an end.

Now Paine had fallen victim to a gigantic counter-revolution in revolutionary disguise which had succeeded in entrenching his original foes -- the monarchy. Hearing from his old friend Thomas Jefferson that he was welcome to return to America he gave up France as a bad job and in September 1802 sailed to Baltimore. His writings bear witness to his revolutionary activities and provide us with a detailed picture of the evolution of social and political change at the end of the eighteenth century.

1.2 The "Rights of Man"

Jonathan Clark, the Hall Distinguished Professor of British History at Kansas University, calls into question the belief that Thomas Paine wrote the whole work in an article on the authorship of *Rights of Man* published in the *Times Literary Supplement* on 16 September 2015. Clark (2015) focuses on a 6000-word long passage, located just before the "Declaration of the Rights of Man and of Citizens" (henceforth, for brevity, the *Declaration*), which forms the central historical passage in Part I. Paine prefaces this passage with the words "I will here, as concisely as I can, trace out the growth of the French revolution, and mark the circumstances that have contributed to produce it" (in Philp, 2008). Clark argues that (i) the passage differs in tone from the rest of the work; (ii) the prose is unlike Paine's; (iii) it is mostly written in a mandarin third person and does not display Paine's typically prominent authorial voice; (iv) the content of the passage is unlike anything else in Paine's writings

since, elsewhere, he showed no significant knowledge of French history and, more tellingly, did not elsewhere reveal knowledge of French high politics. Clark asserts that, examined more closely, its prose seems not to be that of an Englishman at all, reading instead like the English prose of a native French speaker.

Who, then, could have written this 6000-word narrative? Clark suggests that its author was probably The Marquis de Lafayette. It seems to embody not neutral history but his very personal perspective on events and has as hero none other than Lafayette himself. Clark thinks that this passage was expressed in the third person in order to conceal the "vanity" of its author and points out that earlier in the *Rights of Man* Paine acknowledges that Lafayette had provided him with information and at least one document. The strongest piece of evidence that Lafayette had supplied Paine with the substance of the 6000-word narrative comes from Lafayette's own hand since on 12th January 1790 he had written to George Washington "Common Sense [Paine] is writing a Book for you -- there you will see a part of My Adventures."

Clark concludes that this passage is very probably not a history primarily written by Paine but Lafayette's self-serving publicity, part of his attempt to become the "George Washington" of the French Revolution. What better way for Lafayette to publicize his own version of events, with himself at the centre, than feeding his own interpretation to his English friend? The consequences of accepting such an attribution would be to render Paine's important writings open to general re-interpretation. A computer-based stylometric investigation of this attribution is thus merited.

2. Materials: Corpora Collected

For our main study, related to the *Rights of Man*, we collected 106 texts altogether, details of which are given in Appendix 1. We organized this collection in four main groups, as summarized in Table 1. The median size of these 106 text files is 2535 word tokens.

Table 1. Tom-Paine Corpus, summary statistics.

Category	**Text files**	**word tokens**
Controls	45	128217
Marquis	19	45023
Paine	29	161287
Queried	13	18674
Sums:	106	353201

As the works available by the Marquis of Lafayette in English are much fewer than those of Paine, and in genres (letters and memoir) that are absent or rare in Paine's writings we selected works by a number of additional control authors to contrast with Tom Paine's written style. These authors were all contemporaries of Paine, active in the American Revolution (like Paine and Lafayette), with whom he interacted. The texts from these control authors are summarized in Table 2.

Table 2. Breakdown of control authors.

Control authors	**Text files**	**word tokens**
Franklin, B	1	2207
Hamilton, A	19	63384
Jay, J	2	3962
Jefferson, T	12	31812
Madison, J *	11	29168
Sums:	45	128217

[* Four of the "Queried" texts are probably mostly by James Madison as well.]

Having examples from this set of control authors allows us to make comparisons not just between Paine and Lafayette (where the genre distribution is inevitably unbalanced) but between Paine's works and representatives of the kinds of writing on topics of concern to him which were being written and read (by Paine among others) during the time when he was active. The "Queried" texts include the suspect sections of *The Rights of Man* (Paine, 1791), along with several 'distractors', whose role in our analyses will be explained later.

Not all of the text files in our corpus represent individual whole works. Some are sections or chapters of longer works. The 2 cases where a larger work was split into more than 2 files were *Lafayette in America, First Voyage and Campaign* (split into 6 segments of approximately 3000 words each) and the book central to this investigation, Paine's *Rights of Man,* Part I, (Paine, 1791). The latter work (henceforward for brevity referred to as TROM) was broken into 8 portions, as detailed in Table 3.

Table 3. Division of TROM into separate text files.

File name	**description**	**assigned category**	**word tokens**
TROM_Preface	Author's Preface, criticizing Edmund Burke	Paine	821
TROM_Rights_1	Main text from start up to the suspect 6000-word passage	Paine	21460
TROM_Mid1	First approximately-half of suspect passage, beginning "The despotism of Louis XIV., ..."	Queried	2950
TROM_Mid2	Second approximately-half of suspect passage, beginning "For this purpose, the Court set about ..."	Queried	3140
TROM_Declaration	Declaration of the Rights of Man and of Citizens	Queried	828
PaineT_Observations	Passage immediately following the *Declaration*	Paine	711
PaineT_Miscellaneous	Section headed "Miscellaneous Chapter"	Paine	8791
PaineT_Conclusion	Section headed "Conclusion"	Paine	2534

In addition, as part of a subsidiary study on the theme of co-authorship and its stylistic signals, we also collected a parallel corpus, summarized in Table 4, further detailed in Appendix 2. This will be used in subsections 3.4 and 4.1 to study of how well our analytic tools can detect co-authorship, in a naturally-occurring case where we have the great advantage of knowing just where each author's contribution begins and ends.

Table 4. Co-authorship corpus, summary statistics.

Category	**Text files**	**word tokens**
Co-authorships & 'distractor'	6	44837
DH	14	54802
RF	20	71818
Sums:	40	171457

The median size of these texts was 3094 word tokens. Additionally, five individually authored sections of *The Federalist Revisited* (Holmes & Forsyth, 1995) were saved as separate files.

3. Exploratory Analysis using High-Frequency Words

In a pioneering work first published in 1964, Mosteller and Wallace (1984) used relative frequencies of commonly occurring words -- mainly words such as prepositions, conjunctions and articles -- as discriminators to investigate the mystery of the authorship of the *Federalist Papers*. Their scholarly analysis opened the way to the modern, computerized age of stylometry. The use of (mostly non-contextual) high-frequency words as tools in attributional problems was continued by J. F. Burrows (1992) and since then multivariate statistical analyses involving large sets (50-100) of such words have achieved some significant successes. Some noteworthy examples in a wide array of authors and genres have been the attribution of the 1583 *Consolatio*, shown to be not a lost work of Cicero but a sixteenth-century forgery (Forsyth *et al.*, 1999), the investigation into the authorship of the so-called 'Pickett Letters' of the American Civil War (Holmes, Gordon and Wilson, 2001), the identification of the author of the 15th Book of Oz (Binongo, 2003), and the new look at the authorship of the *Book of Mormon* (Jockers *et al.*, 2008). This 'Burrows-style' approach essentially picks the N most common words in the corpus under investigation and computes the normalized occurrence rate of these N words in each text-unit, typically per 100 or per 1000, thus converting a set of P text-units into a P-by-N array of numbers. Multivariate statistical techniques, most commonly Principal Components Analysis and Cluster Analysis, are then applied to the data to look for patterns. The former aims to reduce the dimensionality of the problem by transforming the N variables to a smaller number (usually 2) of new variables and the latter technique provides an independent and objective view of groupings amongst the textual samples by means of a tree-diagram or dendrogram. The Burrows-style approach has become the first port-of-call for attributional problems and will be the initial technique adopted in this investigation.

The choice of text size in stylometric studies is always problematic. Smaller text units are too short to provide opportunities for stylistic habits to operate on the arrangement of internal constituents, while larger units are insufficiently frequent to provide enough examples for reliable statistical inference. Forsyth and Holmes (1996) found the median text block size in a selection of stylometric studies to be around 3500 words. On the other hand, in their study of the *Book of Mormon*, Jockers *et al.* (2008) claim that even the smallest chapters are of adequate size for stylometric analysis, finding no correlation between the correct assignment of an author and the length of text sample. Eder (2015) examined a number of attribution methods in several languages on a variety of text lengths and concluded that 2500 words or more were needed for reliable attribution, though the figure varied somewhat across genres and languages. More recently, Eder (2017) has found that a diversity index can be computed prior to attribution which will indicate whether conditions are such that samples down to 1500 words in length or even slightly smaller can be reliably attributed. This issue is thus still subject to active research, so it was decided in this initial phase of the analysis to start conservatively by working with texts of 2500 words or more in length. This criterion selected 55 text files including the two prime queried texts TROM_Mid1 and TROM_Mid2, of 2950 and 3140 words respectively.

The value of N used varies by application and genre but typically lies between 50 and 100, the implication being that these words should be among the most common in the language and that content words should generally be avoided. Attributional studies have achieved success with N set as low as 50 (Holmes and Forsyth, 1995; Holmes, Robertson and Paez 2001), but Hoover (2004), Eder (2015) and others have found somewhat larger feature sets to be more effective. We selected our features by first generating a list of the most frequent 144 words in our full corpus of 106 texts, then winnowing out any word occurring in fewer than 75 texts (i.e. less than 70% of documents) as well as excluding first-person and second-person and gendered third-person pronouns and possessives. This filtering strategy follows the practice of Hoover (2004).

As it happened, 44 words were removed by this procedure, leaving 100 words to be used as features. These 100 words, as well as the 44 words excluded can be seen in Appendix 3.

Using these words, a grid of size 55 by 100 was produced, with each of the 55 rows representing a text and each of the 100 columns giving the occurrence rates (per hundred) of a particular word in those texts.

3.1 Paine versus the "Founding Fathers"

The first phase in this investigation was designed to assess the validity of the proposed technique. We were interested to discover how clearly the texts of the three "founding fathers" (Hamilton, Jefferson and Madison) that were included in this collection of 55 documents could be separated. More important, we wanted to know if the texts by Tom Paine could be distinguished from those other three authors.

For this purpose, a Principal Components Analysis was performed on 42 rows of this dataset, representing 10 works by Hamilton, 5 by Jefferson, 5 by Madison and 22 unquestionably by Paine. This reduced the 100 variables in the original data to 8 composite variables which between them accounted for 51.74% of the overall variance. Figure 1 shows the 42 texts plotted the space of the first two (most important) of these components (PCs). PC1 accounted for 11.18 percent and PC2 for 9.38 percent of the overall variance (20.56% together).

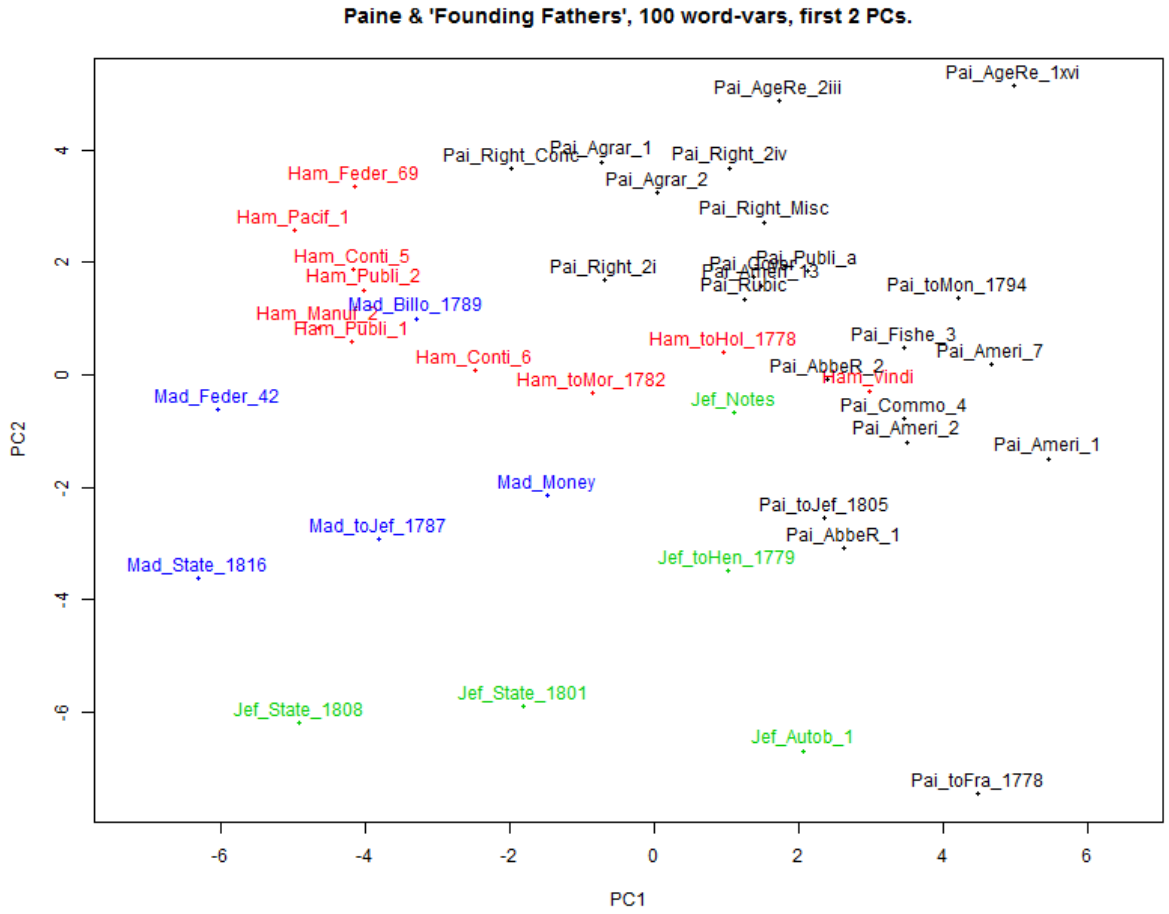

Figure 1. Works by Paine and "founding fathers" plotted in the first 2 components (PCs) of a Principal Component Analysis based on 100 word-rate variables.

This diagram shows an imperfect yet encouraging degree of separation. The first, horizontal, component, clearly expresses a strong authorial signal. Taken together with the second, vertical, component, we find a nearly perfect separation of Paine, in the upper right-hand area of the graph, from the other authors. It also looks as if Hamilton and Jefferson are quite easy to differentiate, mainly on the basis of the vertical dimension.

Broadly speaking, this pattern recapitulates the finding of Sigelman et al. (1997) that Paine's style is quite distinct from the style(s) of his contemporaries, a finding also attested by Berton et al. (2016). However, some caveats are in order. Firstly, one of Madison's texts falls among a group of texts by Hamilton. Given the effort that Mosteller and Wallace (1984) needed to separate these two authors, that is unsurprising. Somewhat more troublesome is the fact that one of Hamilton's longest works, the *Full Vindication of the Measures of Congress*, penetrates the Paine cluster. This is a juvenile work, written when Hamilton was only 17, so we can still argue that Hamilton's mature style can be distinguished from Paine's; but the fact that Hamilton's 1778 letter to Holt also appears on the borderline of Paine's cluster indicates that we must be wary of drawing firm conclusions from such a graph.

The appearance of three State of the Union addresses, by Jefferson and Madison, towards the lower left of the graph suggests that dropping pronominals has not eliminated genre effects, as does the tendency of letters (including part 1 of Paine's rebuttal of the Abbé Raynal, which is framed as a letter) to have low scores on the second component.

To assist in interpreting this graph, the 6 words loading most positively and negatively on these two Principal Components are listed in Table 5. The contrast of "was" with "is" as part of the vertical dimension, PC2, suggests that it taps into differences of genre or stance, thus is not a pure authorial signal. Hence we proceed with caution.

Table 5. Strongest negatively and positively loaded words on first two Principal Components.

Negative loadings on PC1	of, the, well, which, by, its
Positive loadings on PC1	now, were, not, could, but, out
Negative loadings on PC2	on, their, them, at, general, was
Positive loadings on PC2	or, any, upon, it, than, is

3.2 Lafayette versus the "Founding Fathers"

The next step is to see how well our sample of texts by Lafayette can be distinguished from those of the "founding fathers". For this purpose we extracted 30 rows from our 55-item grid -- the same 20 by Hamilton, Jefferson and Madison as in section 3.1, plus the 10 Lafayette samples that are at least 2500 words long.

Another Principal Components Analysis was performed, this time requiring 6 components to account for more than 50% of the overall variance in the data. Figure 2 shows the 30 texts plotted in the space of the first 2 components (PCs). PC1 accounts for 21.48 percent and PC2 for 9.11 percent of the overall variance (30.59% together).

This presents an interesting picture. The samples from *Lafayette in America* are clearly distinct from all the rest, but these seven samples represent sections of only 2 different books, both autobiographical; and the rest of the samples are not well separated by author. Once again Hamilton's *Vindication* stands apart from his more mature writings.

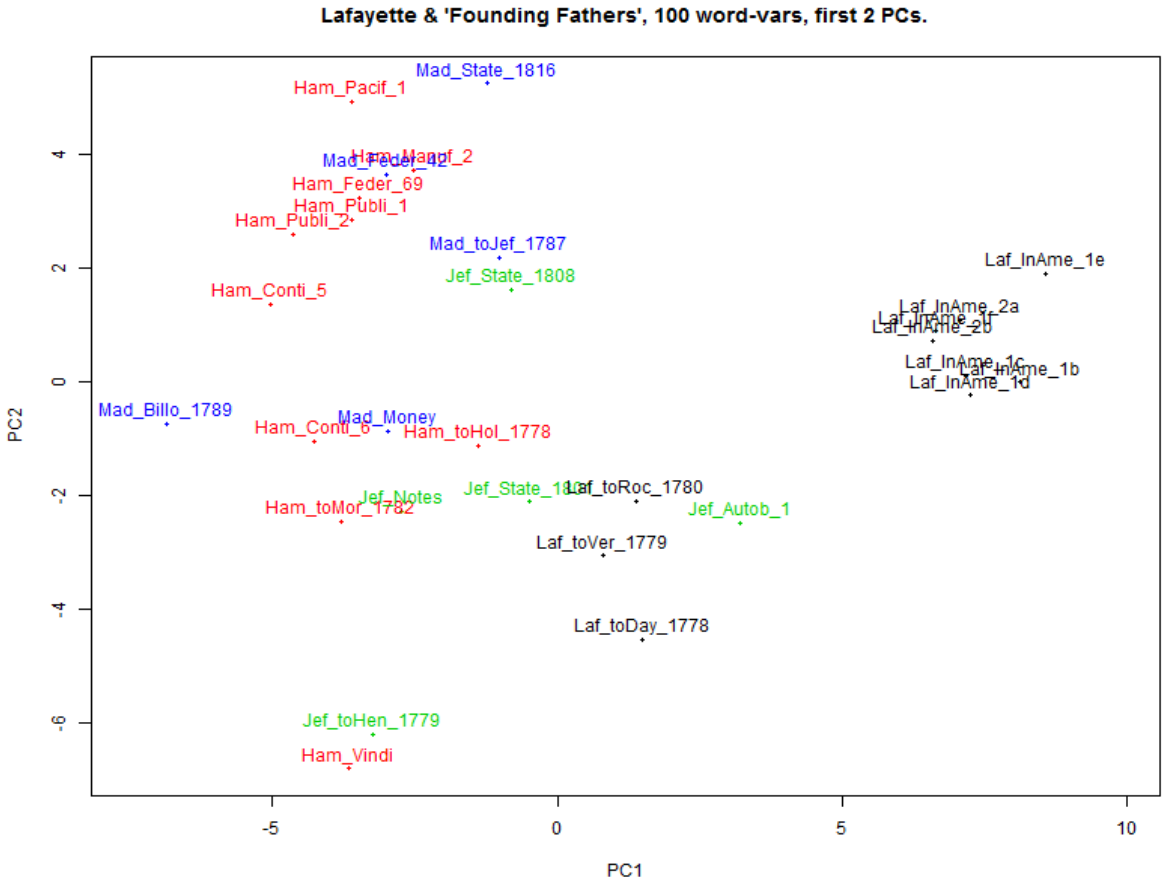

Figure 2. Works by Lafayette and the "founding fathers" plotted in the first 2 components of a Principal Component Analysis based on 100 word-rate variables.

Of particular interest are the three letters by Lafayette, to Dayen, Rochambeau and Vergennes, which are poorly distinguished from Jefferson. These letters were written originally in French and translated into English when Lafayette's collected works were published. We have retained them in our sample out of curiosity. That may be said to introduce noise into the system, but studies by Rybicki (2012) and Forsyth & Lam (2014) have shown, rather counter-intuitively, that authorship attribution is possible in translated works. Moreover, they remind us not to underestimate the potential variability of Lafayette's style if he had written in a wider range of English genres. They can also be seen as proxies for how his style might vary when collaborating with an English-speaking co-author.

Overall then, this part of the analysis shows that *Lafayette in America*, parts I & II are very easy to distinguish from the writings by our sample of the "founding fathers". The three of Lafayette's letters which have, in effect, been edited by another hand differ noticeably from his memoirs, *Lafayette in America*. Despite this, all 10 of Lafayette's samples, including the three translated letters, have positive scores on the first Principal Component, whereas 19 of the 20 texts by the other authors have negative scores on this dimension, the sole exception being Jefferson's *Autobiography*. So here we have a mixed signal which is nevertheless mainly authorial.

The most negatively and positively loaded words on the first two Principal Components are shown in Table 6. The presence of "was" and "were" as opposed to "is" on the dimension that separates Lafayette in America from the rest again suggests a genre effect, while the presence of "general" indicates a topic effect: it arises from the fact that General Washington is frequently mentioned in

descriptions of the Revolutionary campaigns. The strong negative loadings of "they" and "them" on PC2 suggests that even third-person, non-gendered pronouns are affected by other aspects of language than authorship.

Table 6. Strongest negatively and positively loaded words on first two Principal Components.

Negative loadings on PC1	is, or, be, it, as, are
Positive loadings on PC1	two, general, at, were, had, was
Negative loadings on PC2	them, they, could, should, but, do
Positive loadings on PC2	under, in, of, which, an, the

3.3 Lafayette, Paine and the Rights of Man

Thus we have imperfect discrimination, where authorship is to some extent confounded with other signals. Nevertheless Paine in general is relatively well distinguished from his contemporaries and a pair of works by Lafayette is very different from a variety of works by those contemporaries. Even Lafayette's translated letters are atypical when compared to texts by Hamilton, Jefferson and Madison. In short, this approach does offer clues to authorship when applied to writings of the relevant vintage, provided that those clues are taken as evidence rather than proof.

We next examine whether our two principal authors of interest, Lafayette and Paine, are distinguishable; and, in particular, whether the queried parts of TROM resemble one of these authors more than another. To do this we apply PCA to the 10 texts by Lafayette and the 22 by Paine already analyzed in the preceding sections, along with three further samples from TROM.

Figure 3 shows a plot of these 35 texts in the space of the first 2 Principal Components (PCs). PC1 accounts for 20.81 percent and PC2 10.48 percent of the variance in the data (31.29 percent together). Table 7 shows the words with the strongest loadings on these dimensions.

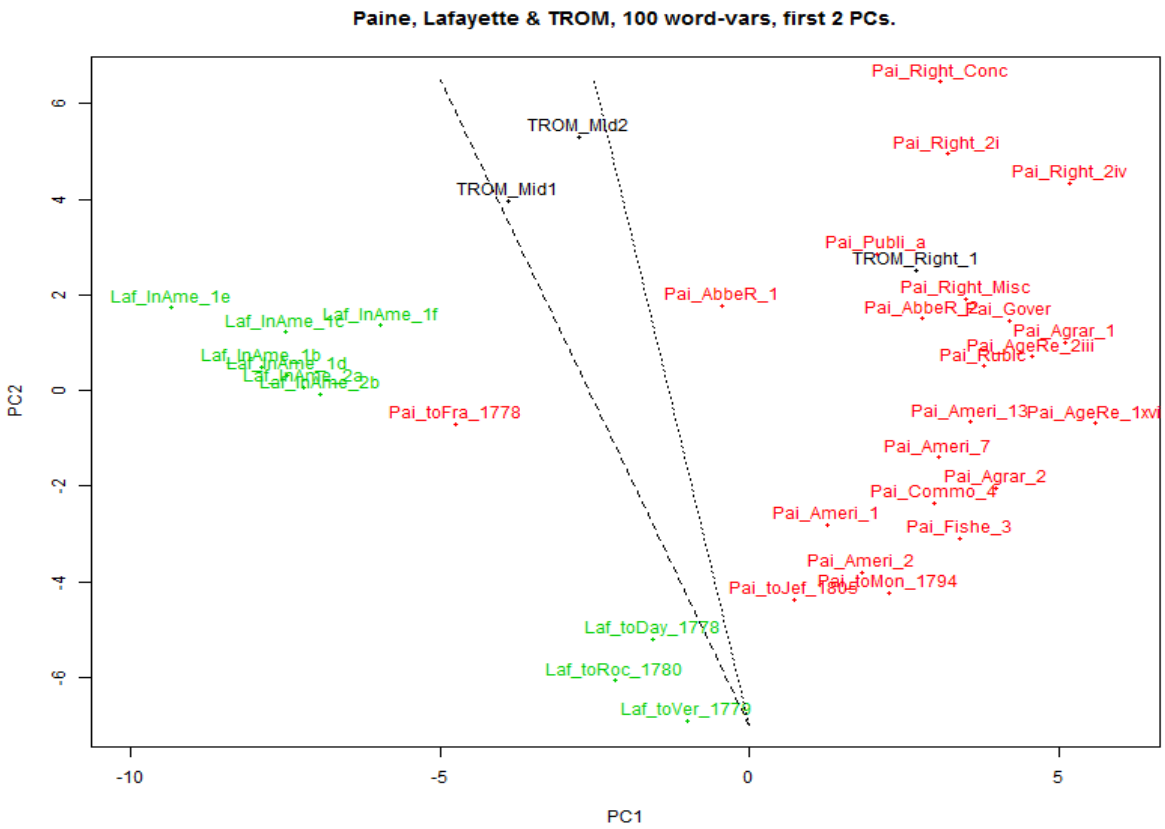

Figure 3. Works by Lafayette and Paine plotted in the first 2 components of a Principal Component Analysis based on 100 word-rate variables.

Again Paine's sample contains an outlier, his letter to Franklin of 1778. Lafayette's texts split into two well-separated groups. Even so, it is possible to separate the undisputed Paine samples from the undisputed Lafayette samples on the first Principal Component with only one exception.

TROM_Rights_1, the first part of *Rights of Man* leading up to the disputed passage falls almost centrally into the Paine cluster. The two parts of the disputed passage, TROM_Mid1 and TROM_Mid2, stand somewhat apart from both authorial groupings, slightly closer to the *Lafayette in America* samples than to the centre of Paine's cluster. To the extent that PC1 represents an authorial dimension, they are more like Lafayette than Paine; but they are not typical of either author. The two dotted lines show that it is possible, with a single exception, to separate Paine's from Lafayette's texts by a straight line -- either with a line that has both queried segments on the same side as Paine or with a line that has both segments on the side of Lafayette.

This graph is thus highly compatible with the assertion that TROM_Rights_1 is solely or mainly by Paine. It is also compatible with the proposition that the queried passage is by neither of them. The idea that it could be a co-authorship is by no means ruled out, an idea that will be followed up in the succeeding section.

Table 7. Strongest negatively and positively loaded words on first two Principal Components.

Negative loadings on PC1	general, was, had, were, two, at
Positive loadings on PC1	any, can, are, it, or, is
Negative loadings on PC2	would, if, have, should, might, make
Positive loadings on PC2	of, each, the, government, its, other

To gain another perspective on the relations among these texts we performed a hierarchical cluster analysis using Ward's method on a distance matrix obtained from the first 6 Principal Components (enough to account for at least half of the overall variance, 53.46% to be precise). Figure 4 shows the result of this clustering.

This clustering presents a twofold grouping at the top level. In the left-hand group are the seven chunks of *Lafayette in America*, tightly bound, with the two parts of the queried passage in TROM as well as two epistolatory texts by Paine linked more distantly. To the right is a group of texts predominantly by Paine. This group also divides rather naturally into two subgroups -- a central group all by Paine including the initial section of TROM and, further to the right, a group by Paine except with the three translated letters of Lafayette intermixed. This supports the idea that TROM_Rights_1 is a typical Paine text. However the two queried passages do not behave unambiguously like texts by either of our two primary authors. The contrast between the two halves of the queried passage (TROM_Mid1 and TROM_Mid2) and the larger section that leads up to it (TROM_Rights_1) is striking.

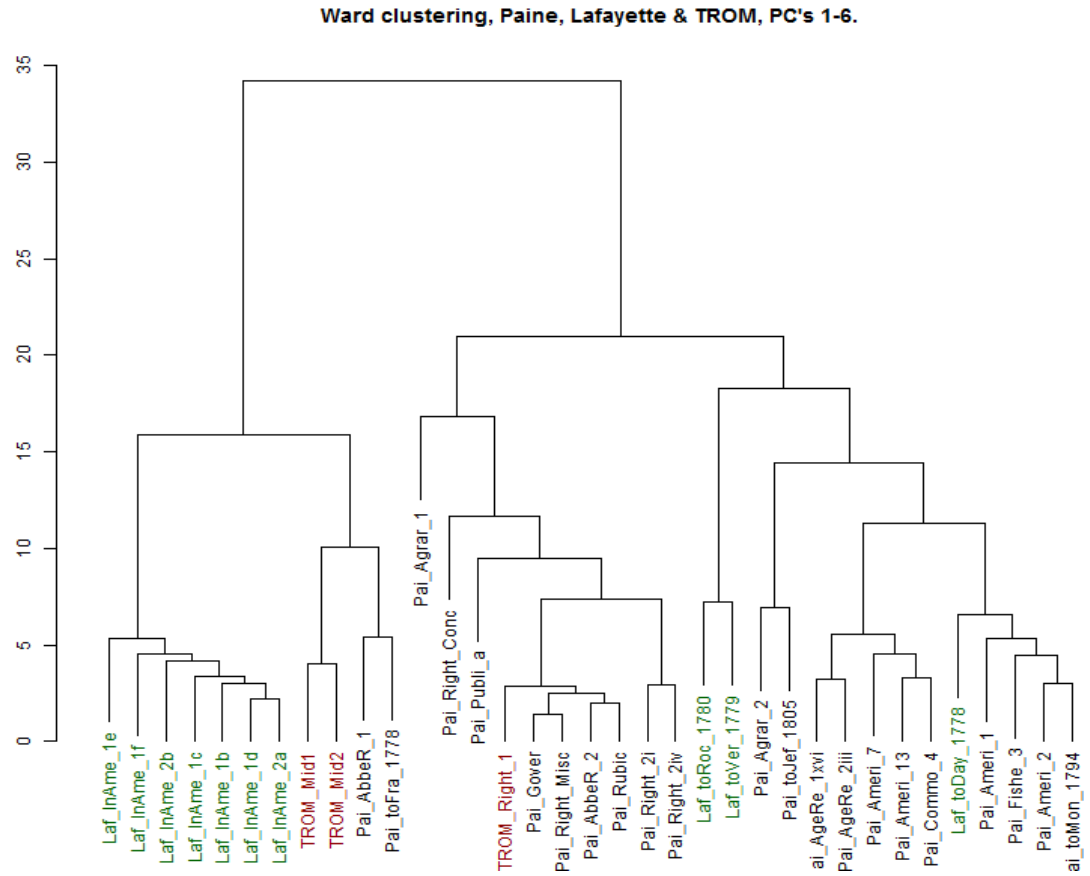

Figure 4. Hierarchical cluster analysis using Ward's method on the first six Principal Components derived from texts by Lafayette and Paine as well as the queried texts.

To summarise this phase of the analysis, it is fair to say that TROM_Rights_1 behaves overall like a typical text by Thomas Paine, while the queried sections do not. There is some evidence linking them to Lafayette, but it is inconclusive. The notion of co-authorship remains plausible.

3.4 Reflexive Co-authorship investigation

To gain a fuller idea of what might be expected in cases of co-authorship, we applied the same approach to the small corpus described in Table 4, above. In other words, we performed a self-examination on texts written by ourselves, including co-authorships, as well as a 'distractor', i.e. a text by another author on the topic of authorship. With this corpus we have the great advantage of knowing who wrote what, in particular who wrote which parts of a particular co-authored text.

Like most researchers, we have been co-authors ourselves on several occasions. We decided to exploit this fact by focussing on a paper with particular relevance to the era of Tom Paine, *The Federalist revisited: new directions in authorship attribution*, which we wrote jointly (Holmes & Forsyth, 1995). Thus this part of our study is effectively a self-examination. We trust that this unseemly degree of self-absorption can be excused by the fact that we have access to a "gold standard" as far as class membership is concerned. In short, we know very well who wrote which sections of the document concerned. This level of certainty is only rarely available in comparable studies, and it provides a background against which to interpret the findings, above, regarding our main focus, Paine and Lafayette.

Co-authorship can take many forms (see: Vickers, 2004; Craig & Kinney, 2009; Rudman, 2016; Fuller & O'Sullivan, 2017) including cases where almost every sentence is the product of more than one hand and the contributors themselves cannot state definitively who wrote which paragraph. Our case, however, is one of the simpler ones. We have divided our *Federalist* article into five segments, very nearly co-extensive with headed sections as published, each of which was written by one author with virtually no interference by the other.

In total, as tabulated in Appendix 2, we collected 45 text files for this subsidiary study, of which 20 were singly authored by Forsyth and 14 by Holmes. In addition, this corpus contains the co-authored text of *The Federalist revisited* as a whole (8979 words) and its five individual subsections, two by Forsyth and three by Holmes. Four further co-authored papers were added, Forsyth and Grabowski (2015), Forsyth & Lam (2014), Holmes & Crofts (2010) and Holmes & Johnson (2012), each jointly written by one of the two authors under scrutiny, with a different co-author. These have file names prefixed with "CO_". Finally, a 'distractor' was added -- pages 1-25 of Chapter 1 by Michael Oakes (Oakes, 2014) on the topic of author identification.

Initially, we performed a Principal Components Analysis, comparable with that in subsection 3.3, on 43 of these files, excluding only the two sections of The Federalist Revisited with lengths less than 800 words. For this purpose 100 words were again selected as features, using the same procedure as described previously in section 3, namely, taking the 144 most frequent words of the corpus and then first removing gendered pronouns (one example) and then 43 other words which were found in fewer than 32 of the 43 documents. The complete list of words, and exclusions, is shown in Appendix 4.

Figure 5 shows these texts plotted in the space of the first 2 Principal Components (PCs) which together accounted for 27.55% of the overall variance. PC1 accounted for 14.92% of the total variance and PC2 for 12.63%.

Here the picture is not clear-cut. The horizontal axis (PC1) does not distinguish the authors. Even knowing the texts concerned it is not simple to interpret this component, although it is influenced by topic: only one of the texts scoring higher than 2.5 on this component relates to linguistic themes, whereas the majority of those scoring lower than this (to the left of the diagram) do relate to linguistic or textual analysis.

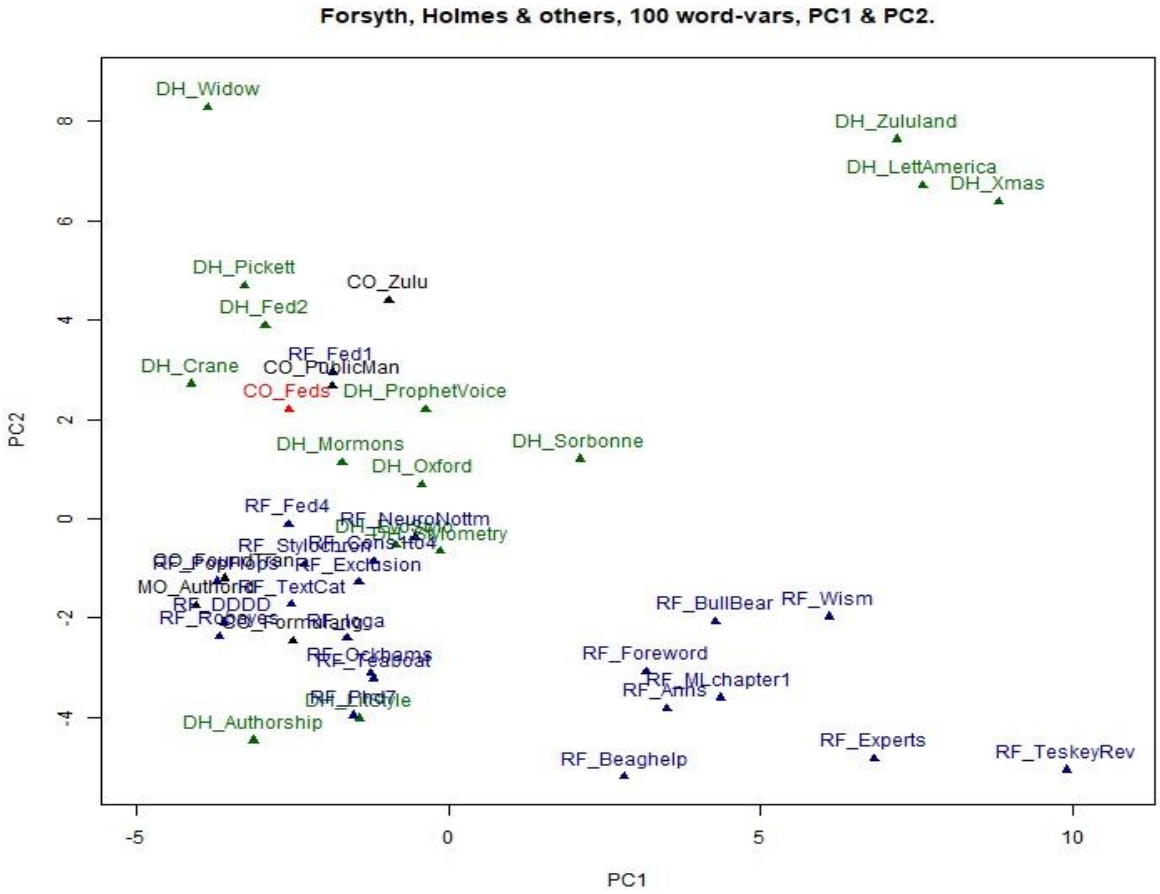

Figure 5. Works by Forsyth (RF), Holmes (DH) and others plotted in the first 2 components of a Principal Component Analysis based on 100 word-rate variables.

However, some degree of authorial discrimination is achieved by PC2. Items in the lower part of the diagram, scoring less than zero on this component, are predominantly by RF, while the majority of points in the upper part are by DH. However, the lower left quadrant is mixed, and rather difficult to view. In fact, a plot of the second and third Principal Components illustrates the situation more clearly, as shown in Figure 6, even though PC3 (which accounts for 7.76% of the total variance) discriminates only weakly between the two main authors. Table 8 shows for each of the first three Principal Components, the five words that load most positively and most negatively.

Table 8. Strongest negatively and positively loaded words on first three Principal Components.

Negative loadings on PC1	used, each, words, by, using
Positive loadings on PC1	so, at, will, all, it
Negative loadings on PC2	is, or, a, be, if
Positive loadings on PC2	were, was, the, and, had
Negative loadings on PC3	on, test, well, here, over
Positive loadings on PC3	analysis, known, their, being, when

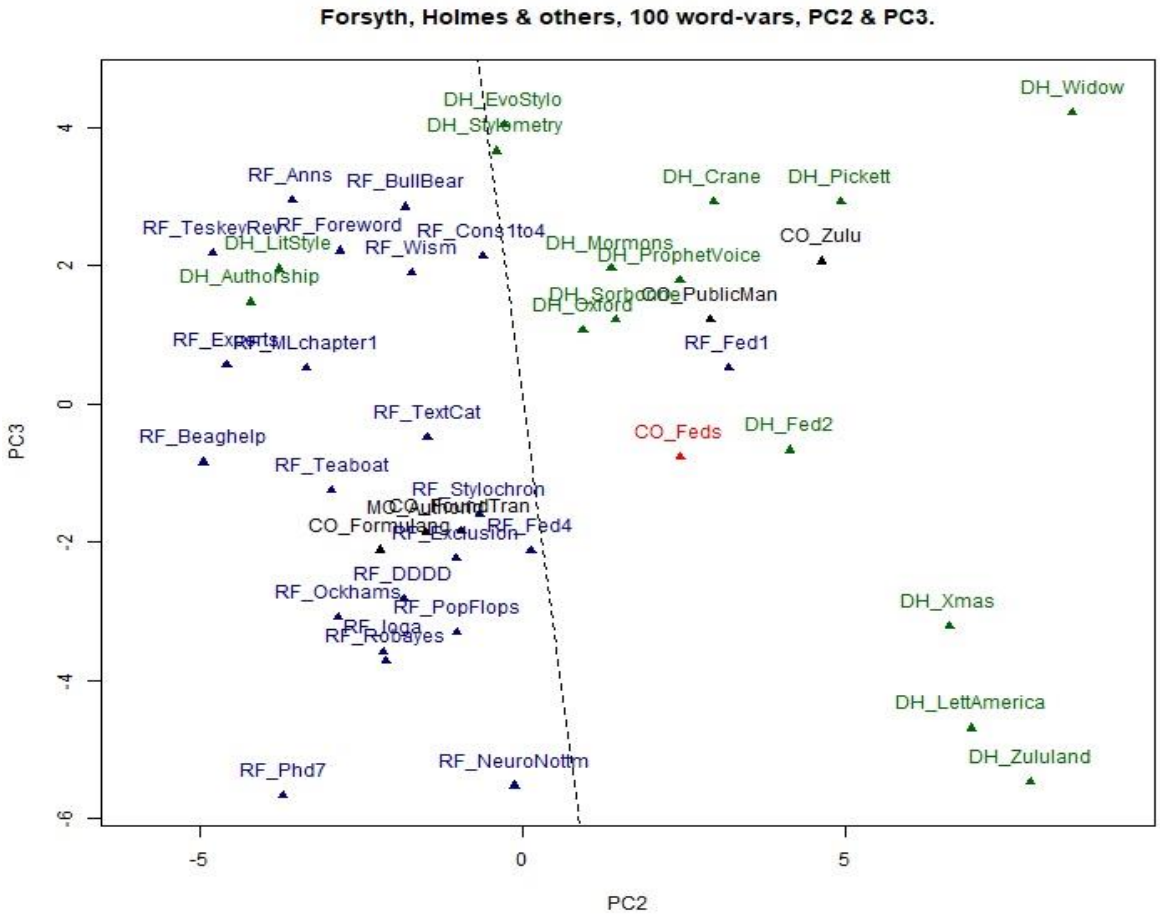

Figure 6. Works by Forsyth (RF), Holmes (DH) and others plotted in the second and third components of a Principal Component Analysis based on 100 word-rate variables.

In Figure 6, PC2, the horizontal axis, still carries the only strong authorship signal, but at least the points are better spread for interpretive purposes. The 34 texts singly authored by either DH or RF (i.e. all except co-authorships, parts of CO_Feds and the distractor, MO_Authorid), can be separated by a straight line, shown dashed in the diagram, with just 2 mistakes. To the left of this line we find all 20 main texts by RF, as well as the two co-authored texts by RF and someone other than DH, and one part (RF_Fed4) of the Federalist paper (CO_Feds). However, we also find two texts written solely by DH and the distractor written by Michael Oakes.

To the right of this dashed line we find 12 of the 14 main texts by DH, as well as DH_Fed2, the longest of the sections of CO_Feds written by DH. We also find the two papers co-authored by DH with someone other than RF. If we were to treat this demarcation line as definitive, however, we would not only assign the whole of CO_Feds to DH, we would also assign part of that paper written by RF (RF_Fed1) to DH as well. It can be seen that while CO_Feds falls near the centre of the DH distribution on PC2, it represents an outlier on that dimension for RF. Interestingly enough, RF_Fed1 is even more of an outlier on that dimension, yet it was written solely by RF.

A conclusion to be drawn from this side study is that while Principal Component Analysis using high-frequency words does very often reveal authorial signals, it cannot on its own be relied on to tease out contributions in a co-authorship. This conclusion is reinforced by Figure 7, which shows the results of a hierarchical cluster analysis, using Ward's method, based on the first 6 Principal

Components derived from this data, which between them account for 51.77% of the overall variance.

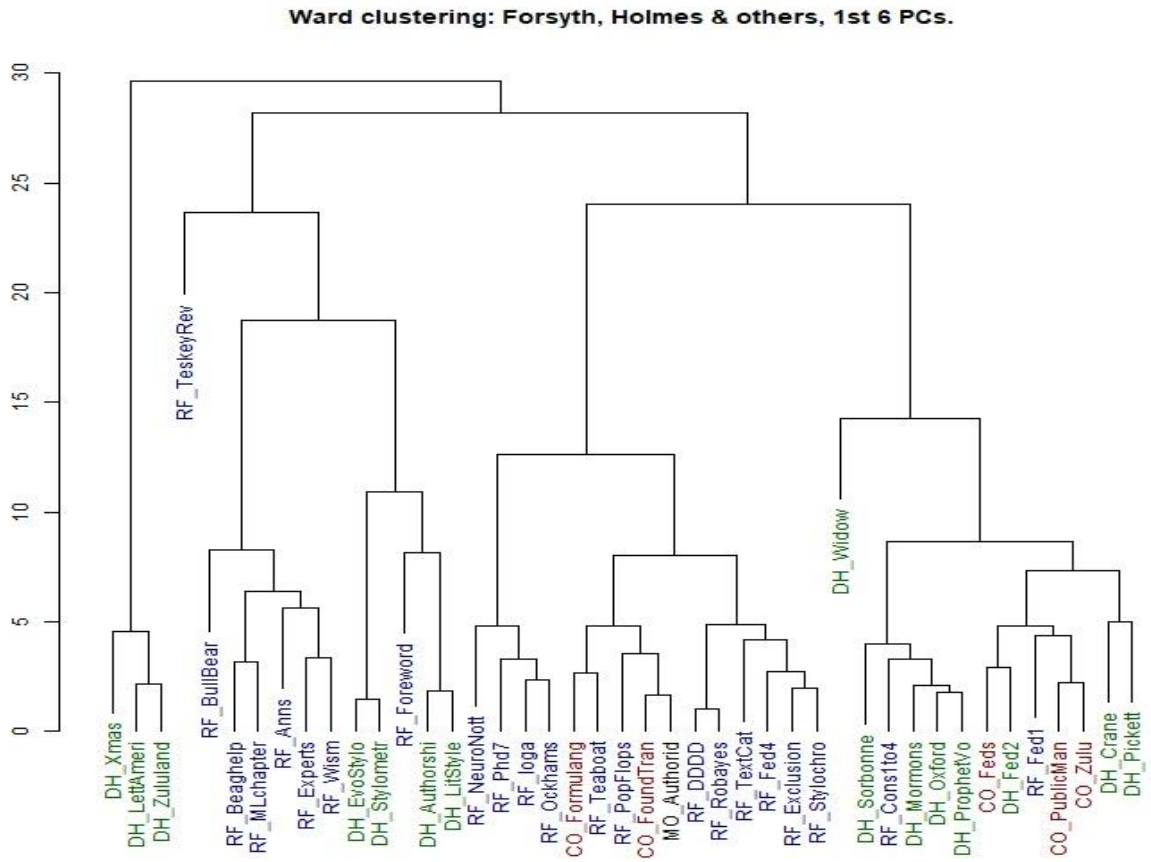

Figure 7. Hierarchical cluster analysis using Ward's method on the first six Principal Components derived from texts by Forsyth (RF), Holmes (DH) and others.

Ignoring the isolate (RF_TeskeyRev) which happens to be the shortest of these texts, we can form five groupings by cutting the vertical lines in Figure 7 horizontally at about level 15 on the vertical axis (though this number has no unambiguous interpretation). The leftmost of these five clusters is a group of 3 DH texts that can be seen as an outlying subgroup in Figures 5 and 6. They are relatively informal writings compared with the rest of the corpus. Next, reading from the left, is a group of 6 texts all by RF on topics other than authorship. Then come five texts of which four are by DH on stylometric themes but one, RF_Foreword, is an introduction by RF to a book on finance. Next from the left comes a large group of 15 texts of which 12 are by RF, 2 are co-authorships by RF and someone other than DH, and one (MO_Authorid) is the distractor written by Michael Oakes. Finally, the rightmost grouping contains 8 texts by DH, 2 co-authorships by DH with someone other than RF and the prime focus text, CO_Feds, written by both DH and RF. But it also contains 2 texts solely by RF (RF_Cons1to4 and RF_Fed1).

On the basis of this phase of the analysis, if we suspected that the paper on the *Federalist Papers* (CO_Feds) contained input from both DH and RF but had no inside information, we could not reject the hypothesis that the whole paper was written by DH, though we might lean towards ascribing part 4 (RF_Fed4) to RF -- rather tentatively. The behaviour of the distractor would, rightly, prevent any more definitive attribution. Of course the distractor had to fall somewhere, so its linkage with RF doesn't invalidate the whole approach; but ideally it would have been an obvious outlier.

Thus it can be said that this classic multivariate approach, based on occurrence rates of high-frequency words, does provide suggestive clues in cases of this type, but cannot be regarded as conclusive. It is, after all, explicitly a mode of exploratory data analysis. To be more definitive, we turn to other, complementary, modes of analysis.

4. Trials with TOCCATA

To gain another perspective on the probable authorship of the queried passages in TROM, we employed the TOCCATA text-classification system, version 9, available at the address below.
http://www.richardsandesforsyth.net/software.html

TOCCATA (which stands for Text-Oriented Computational Classifier, Applicable To Authorship) is designed as a test harness for various text-categorization strategies. It comes supplied with software libraries that enable five different text-classification techniques to be used. In addition, the user may write bespoke libraries (in Python3) to implement other techniques. It allows methods to be assessed by cross-validation on a training sample of texts with known class membership (known authorship in this context) and also applied to classify holdout texts that may include disputed items or classes unseen in the training data.

Experience in this field has taught us that no single method is likely to emerge as "champion" in such exercises, so we decided to apply 6 different methods to the question in hand, one specifically written for this investigation. These methods are briefly outlined in Table 9, and more fully described in Appendix 5. (Unless specifically noted, default values were used for all software options.)

Table 9. Text Categorization Methods Used in TOCCATA.

Method name	**Feature type**	**Inference mode**	**Brief description**
deltoid	frequent words	proximity measured by z-scores	This is essentially Burrows's Delta (Burrows, 2002), but converted to a similarity score from a difference score as 1/D, since TOCCATA works with similarities.
maws	frequent words	"naive" Bayesian probability estimation	This is a "naive Bayesian" word-based classifier based on an adaptation (Forsyth, 1995) of the "robust hand-calculated Bayesian analysis" described in Mosteller & Wallace (1984).
vote	all word tokens	voting	This simple method uses every word-type in the training corpus as a feature. To classify a new text, the frequency of every word in that text is counted; then, for each category, a score is computed by adding a 'vote' to a running total. The vote is sqrt(f) if the token is relatively more common within that category than in the full corpus or else zero. The category yielding the highest score is chosen.
tokspans	spans of frequent words	rank correlation	This method attempts to capture some information inherent in *syllexis*, the co-occurrence of words in close proximity, especially sequential co-occurrence. It starts by finding the most common words in the corpus. As distinctive attributes it uses short spans containing these frequent words. In the experiments reported, the span size was 3, thus such items as ('as', 'well', 'as') or ('the', 'number', 'of') appear as features. Since infrequent words are ignored, tuples of fewer than 3 items are common as

			attributes, such as ('of', 'the'), ('the', 'of') and indeed ('the',) -- the last case indicating a triple containing "the" and two less frequent words.
tagsets	POStag span-sets	rank correlation	This method actually uses the same software module as Tokspans, above, but using sets, not tuples, i.e. with sequence ignored; and with part-of-speech tags (see Appendix 6) not words. The span size used was 5. An example of a feature derived by this procedure is ['dt', 'in', 'nn', 'prp', 'vbd'].
taverns	word n-grams (n=2-4)	proportional non-overlapping text coverage	Taverns borrows a technique from the formulib package, developed to explore formulaic language (available at http://www.richardsandesforsyth.net/software.html). A list of the most frequent n-grams of selected sizes is compiled. Then the similarity of a text to a category is computed as the proportion of tokens in that text covered by any combination of n-grams from the category's high-frequency list.

The first three of these methods use individual words as features. Thus they follow what has become the conventional "bag-of-words" approach, in which texts are treated as collections of tokens, with ordering information completely ignored. However, to quote Z. Harris (1954), "language is not merely a bag of words". Accordingly, we wanted also to attempt to tap into sequential/syntactic information -- at least to some degree -- sequence being essential to language. So half of these six methods, the last three, employ sequences (tokspans & taverns) or co-occurring sets (tagsets) of tokens. As well as using six different feature sets, these methods between them employ five different inferential strategies (tokspans and tagsets being variants of the same procedure).

4.1 The Federalist Re-Revisited

Prior to applying our toolkit of text-classification techniques to the main problem under study, we thought it instructive to discover how the methods performed in a case study where we possess privileged information.

Once again, this preliminary investigation concerns the corpus summarized in Table 4 and detailed in Appendix 2. These texts are those investigated by exploratory methods in subsection 3.4, above, including the co-authored text of *The Federalist revisited* as a whole (8979 words) and its five individual subsections, two by Forsyth and three by Holmes. Four further co-authored papers are included, each jointly written by one of the two authors under scrutiny, with a different co-author. These have file names prefixed with "CO_". Finally, a 'distractor' was added -- pages 1-25 of Chapter 1 by Michael Oakes (Oakes, 2014) on the topic of author identification. For the present study, 29 of these 45 texts were used as training data, leaving 16 as a holdout sample.

The TOCCATA program initially runs a subsampling phase on the training data. In this phase, it repeatedly picks a random subsample of n texts from the training data of N texts, where n = int(sqrt(N)). It then builds a model using the remaining larger sample (size N-n) and uses that model to predict the categories of the items in the smaller sample. In the present case, the larger subsample would contain 24=29-5 texts and the smaller 5 texts. This random subsampling is repeated until the required number of held-out items have been classified (255 trials in the present instance). This implements a mode of cross-validation, meaning that the success-rate statistics printed at the end of this process should be a relatively unbiased estimate of how the method would perform with fresh unseen data.

Part of the output from this phase, using repeated subsampling on the 29 training texts, using the MAWS method, is reproduced below.

```
Confusion matrix :

Truecat =         DH   RF
Predcat : DH     100   32
Predcat : RF       7  116

Kappa value =  0.6929
Precision (%) by category :
DH     75.7576
RF     94.3089
Recall (%) by category :
DH     93.4579
RF     78.3784

cases = 255
cases with unseen category labels = 0
hits =  216
percent hits = 84.71
```

Here the cross-validated success rate is 84.71 percent, respectable though not outstanding in a 2-class problem. We can see from these figures a slight bias, in that Recall is higher (and Precision lower) for DH than RF. This indicates that the method is slightly more likely to predict DH than it should.

More interesting is what happens in the third phase, when the system applies the model generated from all 29 training texts to classifying the 16 holdout texts, which it did not see during the training phase. Here the system classifies each text as belonging to whichever class it estimates to be more likely. It also ranks them according to a certainty measure, called "credence". This is simply the geometric mean of two other scores, labelled "confidence" and "congruity". The system will always assign a class to each holdout case: the point of this ranking is to alert users to cases where this assignment is uncertain.

All TOCCATA models work by assigning similarity scores to each text that measure the similarity of that text to the models of each category in the training data. "Confidence" is calculated from the difference between the maximum of these similarity scores, which determines the category assigned, and the mean of the other similarity scores. A larger difference indicates higher confidence, but the number used in this output isn't the difference itself but an index of how far up in the ranking it would fall among the difference scores obtained during the subsampling phase (of which there were 255 in this example). To be specific, if W is the number of correct decisions with lower difference scores during the subsampling phase and L is the number of incorrect decisions with lower difference scores during that phase, then "confidence" is (W+L/2+0.5) / (S+1), where S is the number of subsampling trials.

The value labelled "congruity" assesses how well the text matches its assigned category, independent of how well it matches other categories. It measures the actual similarity score compared to all similarity scores of the assigned category calculated during the subsampling phase. Congruity is computed as (0.5 + B) / (S+1), where B is the number of cases during the S subsampling trials in which the class selected had a lower similarity score. Thus congruity uses the randomized trials to estimate the empirical strength of similarity of the present case to its assigned category, while confidence estimates how the gap between the chosen category and the rest compares with those encountered during those trials.

Overall, therefore, ranking by credence should give an empirically based indication of how much credibility to attach to each decision on the holdout sample. For the 16 holdout texts under consideration, the output is listed below.

```
==== Posthoc ranking :
rank  credence filename                          pred:true   confidence  congruity
   1    0.9436  DH_ProphetVoice_1991.txt          DH + DH       0.8945   0.9954
   2    0.9185  CO_Zulu_2012.txt                  DH ? CO       0.8555   0.9861
   3    0.8702  RF_Robayes_1996.txt               RF + RF       0.8516   0.8893
   4    0.7525  DH_EvoStylo_1998.txt              DH + DH       0.6758   0.8380
   5    0.6804  DH_Fed2_1995.txt                  DH + DH       0.6211   0.7454
   6    0.6355  RF_Fed4_1995.txt                  RF + RF       0.6172   0.6544
   7    0.6314  DH_Fed3_1995.txt                  DH + DH       0.5703   0.6991
   8    0.5717  CO_PublicMan_2010.txt             DH ? CO       0.5078   0.6435
   9    0.5509  CO_Formulang_2015.txt             RF ? CO       0.5352   0.5671
  10    0.4783  DH_Fed5_1995.txt                  DH + DH       0.4297   0.5324
  11    0.4593  CO_FoundTran_2014.txt             RF ? CO       0.4336   0.4866
  12    0.4593  RF_Stylochron_1999.txt            RF + RF       0.4336   0.4866
  13    0.1955  MO_Authorid_2014.txt              RF ? MO       0.1406   0.2718
  14    0.1941  RF_BullBear_1989.txt              RF + RF       0.1387   0.2718
  15    0.1681  CO_Feds_1995.txt                  RF ? CO       0.1094   0.2584
  16    0.0784  RF_Fed1_1995.txt                  RF + RF       0.0273   0.2248
+?+++++??+?+?+?+
```

Thus, for example, DH_ProphetVoice_1991 with credence 0.9436 is very confidently ascribed to DH; while RF_Fed1_1995 with credence 0.0784 is ascribed to RF but with very little confidence. It is correct, but only just. (In this example all 10 holdout texts of known authorship are correctly classified.)

The above listing illustrates the use of a single method, MAWS. Table 10, below, collates the results of six such runs. The 16 holdout texts are listed in Table 10 along with their rankings when each of the 6 chosen methods was used to create a model on the training data and that model applied to the holdout texts.

Table 10. Holdout texts ranked on the DH-RF polarity, consensus of 6 methods.

filename	**total**	**deltoid**	**maws**	**vote**	**tokspans**	**tagsets**	**taverns**
DH_ProphetVoice_1991.txt	87	16	16	12	14	14	15
CO_Zulu_2012.txt	78	14	15	13	13	15	8
DH_EvoStylo_1998.txt	65	10	13	15	10	8	9
CO_PublicMan_2010.txt	61	15	9	5	16	11	5
DH_Fed2_1995.txt	46	7	12	11	5	-2	13
DH_Fed3_1995.txt	35	1	10	8	-1	3	14
DH_Fed5_1995.txt	12	3	7	1	-2	5	-2
RF_Fed1_1995.txt	9.5	6	-1	-2.5	3	1	3
CO_Feds_1995.txt	-6.5	8	-2	-2.5	-7	-4	1
MO_Authorid_2014.txt	-36	-2	-4	-4	-9	-10	-7
RF_BullBear_1989.txt	-38	-4	-3	-10	-6	-9	-6
RF_Stylochron_1999.txt	-45.5	-9	-5.5	-9	-4	-6	-12
CO_FoundTran_2014.txt	-51.5	-5	-5.5	-6	-12	-13	-10
RF_Fed4_1995.txt	-55	-11	-11	-14	-8	-7	-4
CO_Formulang_2015.txt	-65	-12	-8	-7	-15	-12	-11
RF_Robayes_1996.txt	-86	-13	-14	-16	-11	-16	-16

The entries in this table are ranked according to the values in the column labelled "total". This is simply the sum of the values in the last 6 columns. Each column contains the rank position of the text concerned when a model of the method named, trained on the 29 training documents, is applied to this holdout sample. Because this problem has been framed as a dichotomy (DH versus RF) the rank position in the output has been signed, positive if the class assigned is DH and negative if it is RF. Sixteen texts were under consideration, so 16 means a very certain decision for DH, -16 a very certain decision for RF; while 1 or -1 would mean that the decision was ranked last on creditworthiness; and so on.

Overall, this aggregation of results from 6 methods shows that DH_ProphetVoice_1991 was judged most surely by DH and RF_Robayes_1996 most surely by RF. Of the top 6, four are by DH and two by DH with a co-author not present in the training sample. Of the bottom 6, four are by RF and two, again, by RF with co-authors absent from the training sample.

The interesting items are the middle four, i.e. those decisions where the system is most uncertain. Two of these are portions of *The Federalist revisited;* a third is the full paper itself, and the fourth is the 'distractor' by Michael Oakes (also absent from the training sample), on author identification.

This output is more naturally interpreted with the aid of a graph. Figure 8 shows this data in 2 dimensions. The horizontal axis simply shows the scores used to rank Table 4b divided by 96, its theoretical maximum. The vertical axis is the average credence score given by TOCCATA to the decision. It is an index of certainty, as explained above.

Since these scores are related, an approximate U-shape or V-shape of the plotted points is inevitable. However, the 'valley' of the V is useful in identifying more clearly than in a table the dubious texts, those that do not belong securely at either end of the polarity. Dotted lines demarcate this area.

This gives us an idea of what to expect when the same procedure is applied to a more genuinely problematic case. It is apparent that the document written jointly by both the contrasted authors (CO_Feds_1995) sits almost exactly on the borderline of the polarity between them. Interestingly, we see that co-authored pieces where the second author isn't part of the contrast in focus can be assigned quite confidently to the author who is part of that contrast. The single distractor, MO_Autho_2014, does not gravitate strongly towards either pole, although appearing somewhat more Forsythian than Holmesian. The two pieces that land so close together as to render them illegible are CO_FoundTran_2014 and RF_Fed4, both on the RF side.

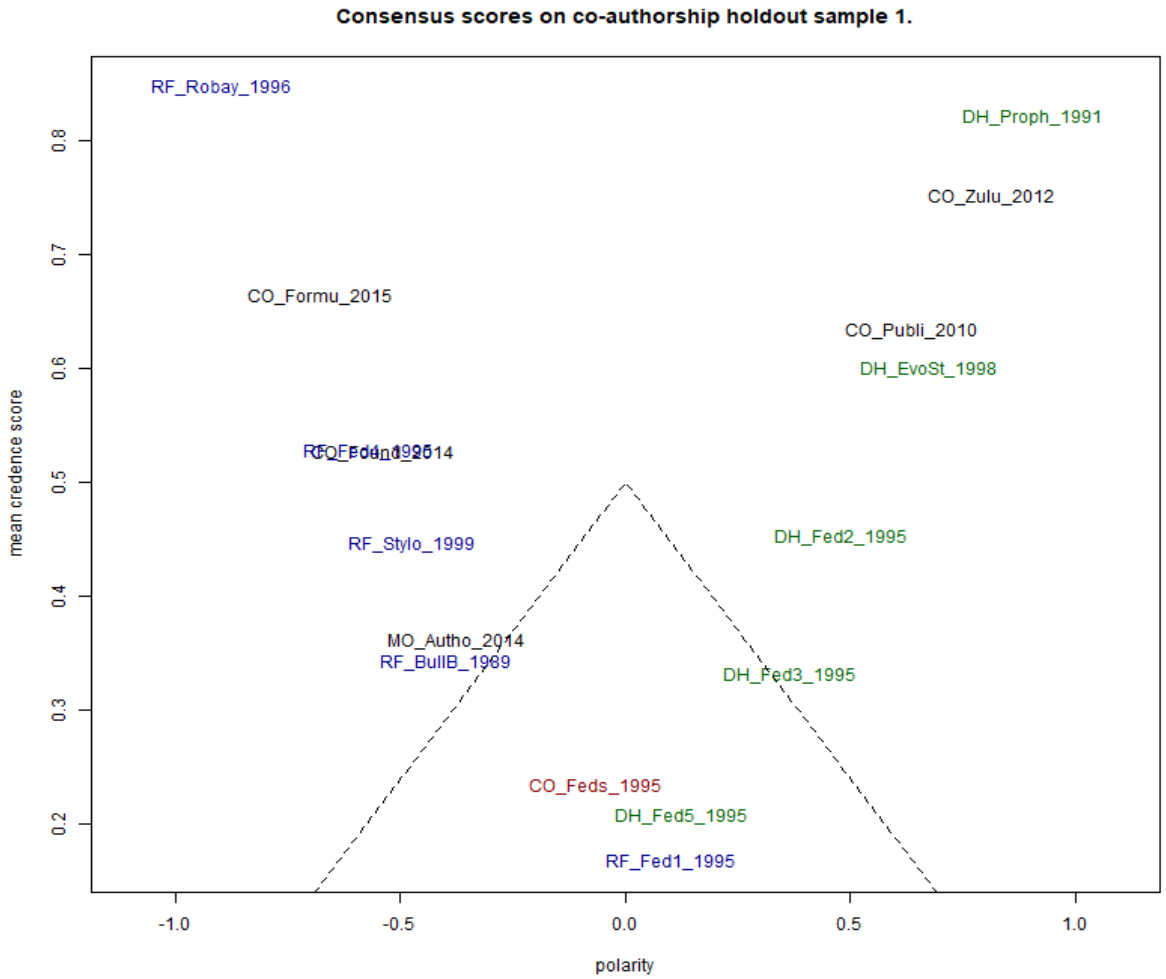

Figure 8. Holdout co-authorship texts plotted according to credence (vertical) and direction (horizontal).

From this graph it would seem that anything with a mean credence score of less than about 0.4 cannot confidently be ascribed to either of the two authors being compared, as this implies a relatively neutral polarity whose small difference from zero could easily be due to chance. This would put 6 of the 16 texts into a zone of uncertainty. The two texts deepest into this zone of uncertainty are RF_Fed1 and DH_Fed5. These are the (historical) introduction and conclusions sections respectively. Although we wrote these individually, they are the most general sections, arising out of many discussions and attempting to convey a coherent overall message. Perhaps it is unsurprising that we achieved something like a common style for this joint enterprise. By contrast, in parts 2, 3 and 4, we were describing the results of applying our own specialized analytical techniques. It is evident, therefore, that despite using methods that rely mainly on frequent function words or part-of-speech tags, the factors of topic and/or genre still have some influence on the classification process. Hence this method is only partially immune to effects other than authorship. [See endnote 1.]

Thus, turning towards Paine and contemporaries, it must be acknowledged that we have no magic bullet, but we do have a way separating some of the stronger signals from the noise.

4.2 Paine and Contemporaries

For the corresponding experiment on the *Rights of Man*, we used the 106 text files detailed in Appendix 1. The same six methods (Table 9) were applied using the TOCCATA software. This time, 20

texts were chosen as a holdout sample, using the other 86 as training data. The holdout texts are listed in Table 12, along with their (signed) rankings, derived as explained in the previous section. Again, the task for the classifier was framed as a dichotomy, to distinguish Paine from all other authors in the corpus.

The idea of treating Non-Paine as distinct text category raises several issues, not least of which is: how could anyone sample the writings of this composite 'author'? This choice is primarily dictated by the practical consideration that we have many fewer documents in English by the prime suspect for the other contributor to TROM, namely the Marquis of Lafayette, than by Paine -- and those only in genres (autobiography and letters) which are relatively rare in Paine's corpus. Nevertheless, we find encouragement in taking this route from the finding of Sigelman et al. (1997) that Paine's writing style is indeed markedly different from 14 other authors contemporary with him. Thus, posing a contrast of Paine versus the rest (where the rest includes Lafayette among seven other authors: see Appendix 1) is not purely fanciful. The ultimate justification is pragmatic: does it shed light on the question at stake?

As illustration, an extract from the output listing using tokspans method with spansize of 3 is reproduced below.

```
==== Posthoc ranking :
rank  credence filename                         pred:true       confidence  congruity
   1    0.9130  TROM_Rights_1.txt                Paine + Paine       0.8397   0.9926
   2    0.8424  Queried_Federalist_19.txt    NonPaine + NonPaine     0.7710   0.9205
   3    0.8120  PaineT_Government.txt            Paine + Paine       0.7061   0.9338
   4    0.7883  PaineT_Rubicon.txt               Paine + Paine       0.6870   0.9044
   5    0.7529  PaineT_AgeReason_2iii.txt        Paine + Paine       0.7634   0.7426
   6    0.6525  Lafayette_InAmerica_2b.txt   NonPaine + NonPaine     0.7252   0.5872
   7    0.5705  Queried_Federalist_20.txt    NonPaine + NonPaine     0.4332   0.7513
   8    0.5462  Queried_Federalist_57.txt    NonPaine + NonPaine     0.5363   0.5564
   9    0.5348  JohnsonS_Idler103_1760.txt   NonPaine + NonPaine     0.5840   0.4897
  10    0.4821  PaineT_AbbeRaynal_1.txt          Paine + Paine       0.2748   0.8456
  11    0.4783  Queried_Federalist_52.txt    NonPaine + NonPaine     0.6069   0.3769
  12    0.3063  TROM_Mid1.txt                    Paine ? Queried     0.3645   0.2574
  13    0.2893  JeffersonT_NotesVirginia.t   NonPaine + NonPaine     0.0992   0.8436
  14    0.2823  Lafayette_toWashington_177   NonPaine + NonPaine     0.3416   0.2333
  15    0.1829  TROM_Declaration.txt         NonPaine ? Queried      0.2557   0.1308
  16    0.1689  Queried_AfricanSlavery.txt   NonPaine ? Queried      0.2099   0.1359
  17    0.1479  TROM_Mid2.txt                    Paine ? Queried     0.1565   0.1397
  18    0.1437  Queried_UnhappyMarriages.t   NonPaine ? Queried      0.1202   0.1718
  19    0.0293  Shax_sonn109.txt                 Paine - NonPaine    0.1164   0.0074
  20    0.0202  Theo_749f.txt                    Paine - NonPaine    0.0553   0.0074
++++++++++?++????--
```

Once again the credence ranking does what it is meant to do: there are 2 mistakes, but they lie at the bottom of the ranking; in fact, the last 6 of these 20 items consist of four queried cases and two errors, both the latter by authors outside the training data -- the final item even in another language.

The top 10 NonPaine & top 10 Paine token spans actually chosen in this run are shown in Table 11.

Table 11. Distinctive NonPaine & Paine token 3-spans.

Most discriminatory **NonPaine** token spans	('of', 'the') ('the',) ('to',) ('to', 'the') ('of',) ('the', 'of') ('by', 'the') () ('their',) ('in', 'the')
Most discriminatory **Paine** token spans	('is',) ('and',) ('and', 'the') ('it',) ('it', 'is') ('or',) ('they',) ('of', 'a') ('as',) ('has',)

This shows that many of the items are shorter than the maximum spansize of 3. For example, the span ('it', 'is') defines a triple containing 'it' and 'is' in that order along with one other word not in the high-frequency list, which in this case consisted of 112 words. Thus "perhaps it is", "it surely is" and "it is unlikely" would all count as examples of this feature.

The results of aggregating the six selected methods are summarized in Table 12.

Table 12. Holdout texts ranked on Paine-NonPaine polarity, consensus of 6 methods.

filename	total	deltoid	maws	vote	tokspans	tagsets	taverns
Queried_Federalist_19	96	17	17	13	19	15	15
Queried_Federalist_52	94	14	16	16	10	18	20
Queried_Federalist_57	84	11	11	17	13	13	19
Lafayette_InAmerica_2b	82	13	10	20	15	10	14
Queried_Federalist_20	69	16	8	10	14	12	9
TROM_Declaration	44	5	15	3	6	9	6
Lafayette_toWashington_17790613	40	9	9	8	7	3	4
JeffersonT_NotesVirginia	38	10	-2	2	8	8	12
Theo_749f	30	4	20	5	-1	1	1
JohnsonS_Idler103_1760	5	-2	1	-7	12	-7	8
Shax_sonn109	4	1	18	-18	-2	2	3
TROM_Mid1	-7	8	5	-9	-9	5	-7
Queried_AfricanSlavery	-17	-3	-6	-15	5	4	-2
Queried_UnhappyMarriages	-18.5	-6	-4	-11	3	-6	5
TROM_Mid2	-42.5	-7	-4	-4	-4	-14	-10
PaineT_Government	-58	-19	-14	1	-18	-19	11
PaineT_AbbeRaynal_1	-69	-15	-7	-6	-11	-17	-13
PaineT_Rubicon	-89	-18	-13	-12	-17	-11	-18
PaineT_AgeReason_2iii	-92	-12	-12	-19	-16	-16	-17
TROM_Rights_1	-109	-20	-19	-14	-20	-20	-16

This holdout sample includes five texts treated as being uncontentiously ascribed to Paine (including TROM_Rights_1 which is the initial part of the *Rights of Man*, roughly 21400 words in length, leading up to the queried passage). There are also four dubious texts. AfricanSlavery and UnhappyMarriages have been de-attributed by scholars working at The Institute for Thomas Paine Studies at Iona College:
http://thomaspaine.org/pages/writings.html#deattributed

The two other dubious texts are TROM_Mid1 and TROM_Mid2, the first and second half of the 6000-word passage immediately preceding the *Declaration* identified by Clark (2015) as likely to be by Lafayette. (See Table 3.) The other 11 texts are unambiguously not by Paine, including the *Declaration* itself, which is identified in Paine's book as a quotation. The four *Federalist* essays are only "queried" in the sense that James Madison probably wrote most of them, but Alexander Hamilton may have contributed to some extent (mainly to numbers 19 and 20). In any case, they definitely weren't written by Paine.

Finally, the non-Paine texts include three 'distractors' which are by authors absent from the 86 training texts. One is William Shakespeare's sonnet 109 (at a mere 117 words easily the shortest in the whole corpus); another is an Idler essay by Samuel Johnson; the third is a letter by Theo Van Gogh to his brother Vincent, written in March 1889, which is in French! We believe that subjecting such 'outsiders' to the same analytic techniques as more directly relevant documents provides valuable clues regarding the interpretation of the results, which are portrayed graphically in Figure 9.

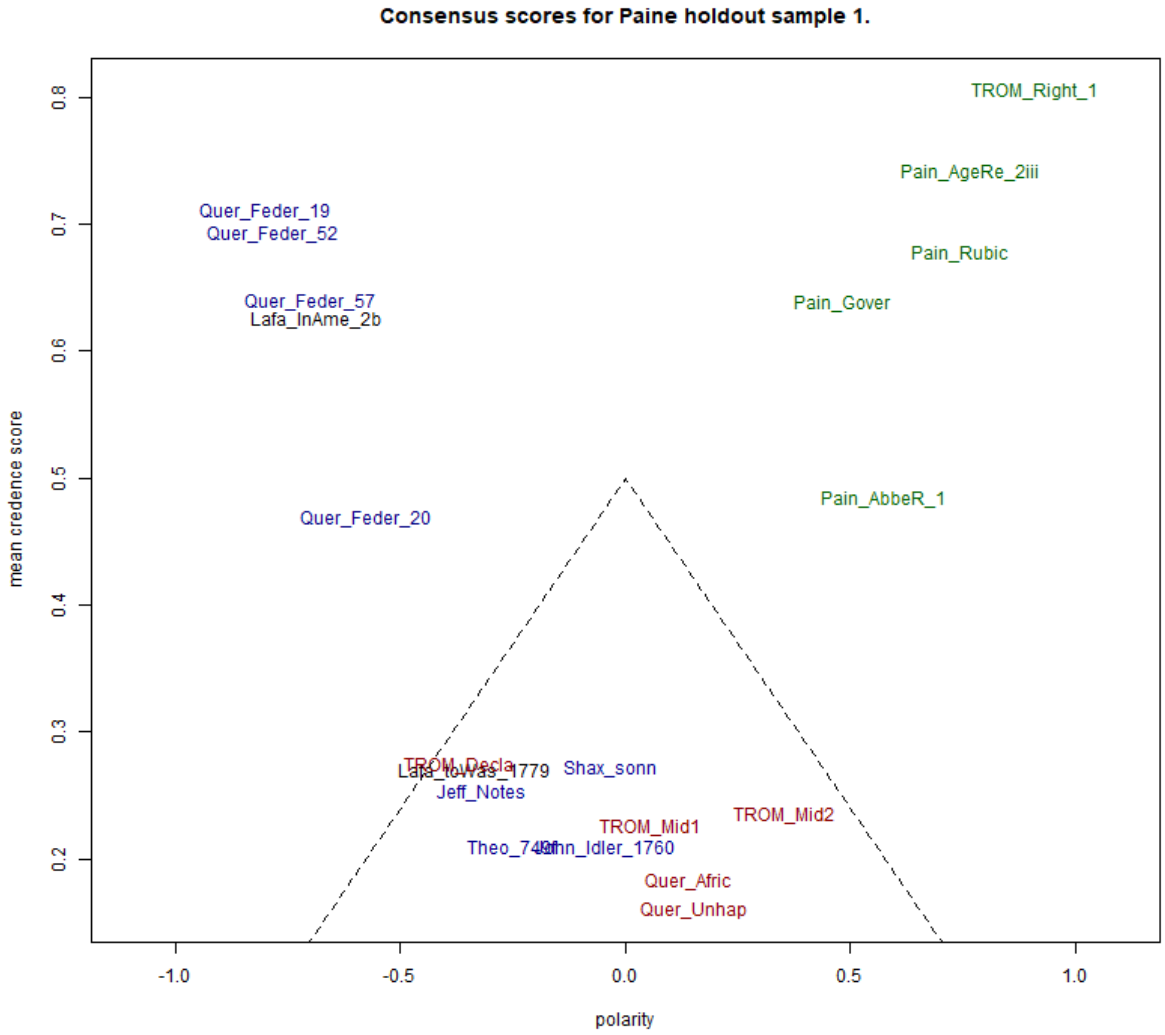

Figure 9. Paine/non-Paine holdout texts plotted according to credence (vertical) and polarity (horizontal).

In this case, all five texts unambiguously by Paine are clearly shown as such. In addition the four *Federalist* papers (Quer_Feder_19, _20, _52 and _57) and the extract from *Lafayette in America*, Part 2, are clearly marked as unlike Paine. Thus all 10 items with a credit score above 0.4 are correctly attributed, five to each side. Two of the non-Paine texts with authors present in the training data, namely Lafayette's letter to George Washington (Lafa_toWas_1779), written in June 1779, and Jefferson's Notes on the State of Virginia (Jeff_Notes) fall on the non-Paine side of the polarity, but with relatively weak confidence. In the case of Lafayette's letter, it is interesting to note that it falls almost exactly in the same position as TROM_Declaration, i.e. the *Declaration*, English translation. Lafayette is known to have played a part in drafting this famous declaration, so the idea that he assisted Paine with rendering it into English is very plausible.

The three clear-cut 'distractors', by Shakespeare, Johnson and Theo van Gogh, all fall deep into the zone of uncertainty, across the half-way point away from Paine, though not by much. Here the method is performing exactly as required. As regards the genuinely queried items, all four of them also fall into the zone of uncertainty. On this evidence, there is no reason to doubt Iona College's de-

attribution of both *African Slavery in America* and *Reflections on Unhappy Marriages* (Quer_Afric, Quer_Unhap). They certainly do not typify Paine's writing style. Nor do the two halves of the suspect 6000-word section identified by Clark. Given that TROM_Rights_1, the part of the book leading up to them, emerges as the most strongly Paine-like of all 20 holdout texts, this corroborates the contention that the queried passage is unlikely to have been written by Paine alone. In fact, these two pieces behave under this form of analysis almost exactly like the unquestioned co-authorship (by Holmes & Forsyth) of section 4.1, falling near the borderline on this polarity between non-Paine and Paine.

We repeated this procedure a number of times, selecting different training and holdout sets at random, though keeping *African Slavery in America, Reflections on Unhappy Marriages*, TROM_Rights_1, TROM_Mid1 and TROM_Mid2 always in the holdout sample. The results were essentially similar, as exemplified by the plot in Figure 10.

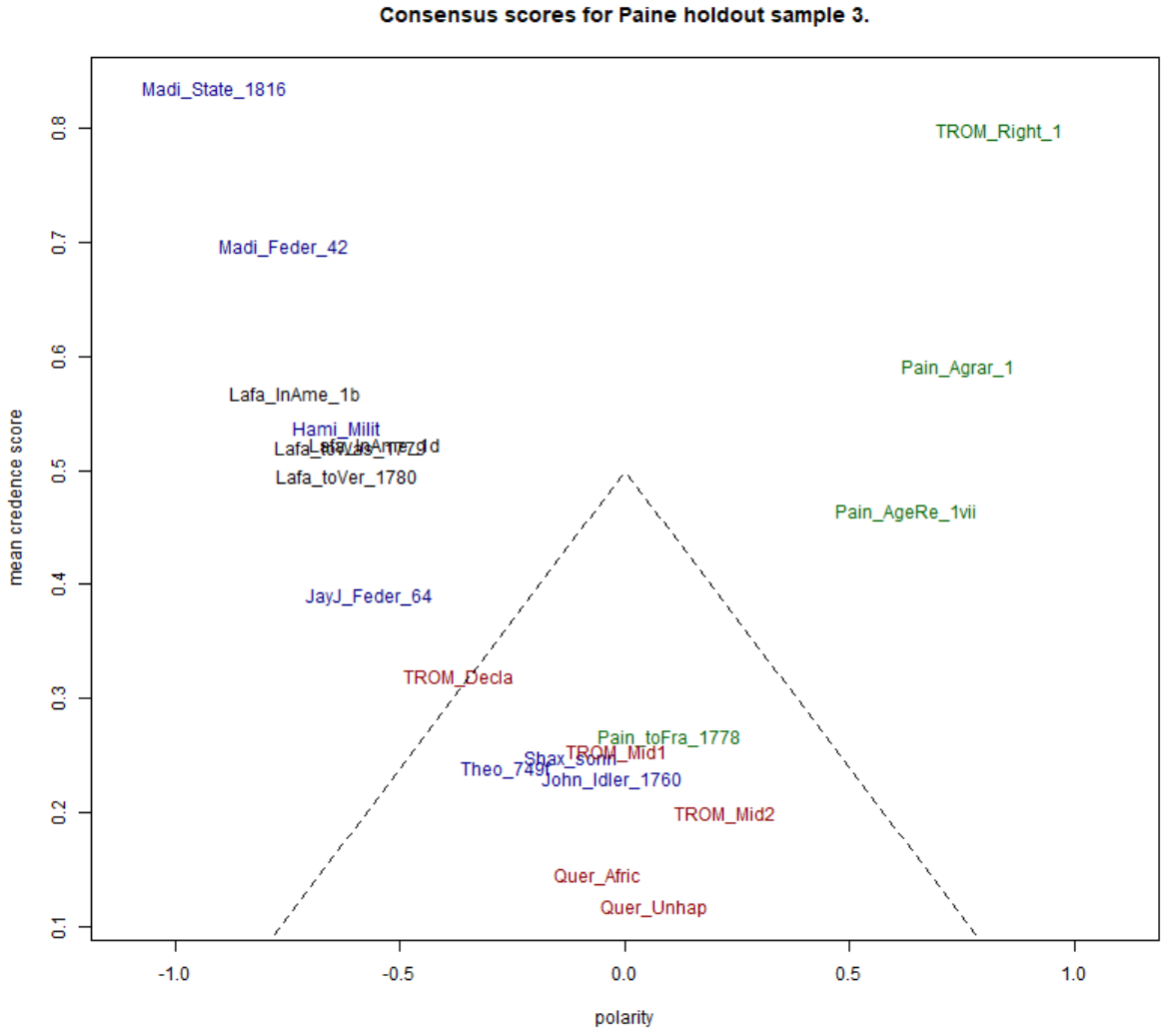

Figure 10. Paine/non-Paine holdout texts plotted according to credence and polarity, alternative training/holdout split.

We regard these results as further supportive evidence for Jonathan Clark's contention. In this connection it may be worth noting that TROM_Mid2, the second half of the doubtful passage is rated somewhat more similar to Paine than the first half. This could be a genre effect, in that the

former section is a brief history of the dramatic events leading up to the King's journey from Versailles to attend the Parliament in Paris in which a starring role is played by a certain M. de la Fayette -- rather like some parts of the autobiographical work *Lafayette in America*. TROM_Mid2 also deals with stirring events, including the storming of the Bastille, but this mainly consists of a reference to a fuller description earlier in the book, and in general this part is rather more in keeping with the reflective parts of the book which come after the *Declaration*, and are presumably by Paine himself. It is not unreasonable to suppose that, even if Lafayette provided a ready-made account, Paine might more heavily edit the more philosophic portions than accounts of events to which Lafayette was personally a witness. Still, this difference is minor: neither portion is typical of Paine's writing.

In summary, this form of analysis, based on using the TOCCATA package, tends to support the findings of our analysis in section 3, based on a Burrows-style approach. This lends further weight to the assertion that the queried 6000-word portion of the *Rights of Man* was probably not written solely by Thomas Paine.

5. "Rolling" Stylometry

Up to this point, we have used prior knowledge, in the case of our own paper, or prior hypothesis, in the case of TROM, to divide our questioned texts into portions for individual scrutiny. This approach could be viewed as falling within a hypothesis-testing paradigm. Recently, however, an alternative, more exploratory, mode, known as "rolling stylometry", has been developed by a number of researchers. (See, for example: van Dalen-Oskam et al., 2007; Burrows, 2010; Craig & Burrows, 2012; Rybicki et al., 2014; Eder, 2016).

In this approach, a questioned text is divided sequentially into fixed-length blocks, usually overlapping, and each block is compared to a training sample of texts with known authorship (or, more generally, known class membership). Then some measure of proximity or distance to the known texts (or an aggregation of them by category) is computed for each block, and these scores are plotted as y-values against sequential position in the questioned text.

This approach suits our research question very well, so we decided to apply the rolling-stylometry functions of the stylo package (Eder et al., 2016) to the full text of TROM and plot the results.

To produce Figure 11, we selected 14 of our larger texts, nine by Paine, four by Lafayette and one by Hamilton, for comparison with TROM and used the rolling.delta() function of the stylo package. (For this purpose we joined together the separate portions of "*Lafayette In America*" I & II, to restore them as two whole works; we also joined the two parts of Paine's pamphlet entitled "*Letter to the Abbé Raynal*".) The x-axis gives the starting position of each block within the 40000+ words of TROM. The y-axis uses Burrows's Delta as a distance measure, based on the 100 most frequent word tokens. Block size chosen was 1000 words with an increment of 500; thus the first block consists of words 1-1000, the second of words 501-1500, the next of words 1001-2000 and so on.

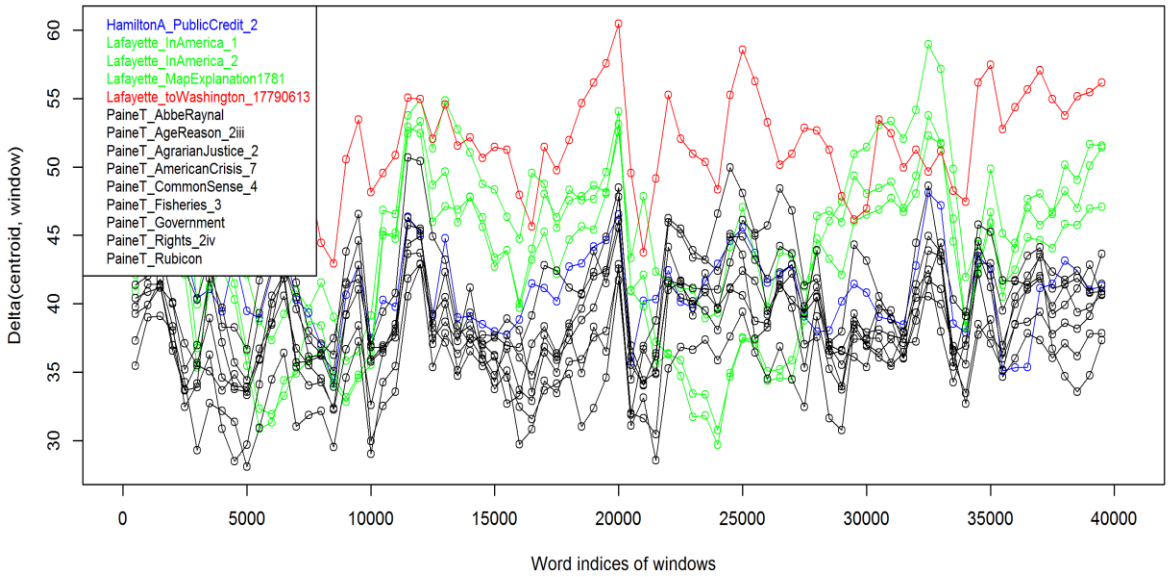

Figure 11. Rolling Delta scores of successive 1000-word blocks of TROM compared to 14 other texts.

The black points and lines show the Delta-distances of each block from the texts by Paine; the green and red lines show distances to Lafayette's works; and the blue line, for reference, shows distances to the control text by Hamilton. The red line is in almost all places the most distant from the TROM sections, presumably indicating a genre difference: it is an inter-personal letter (written in English). However, there are three sections where the lowest, i.e. most similar, text is by Lafayette. The first of these starts at position 6000; the second at position 9000; and the third falls between 23000 and 25000, almost in the middle of the passage queried by Clark.

We take this third section as further corroboration of the results reported above in sections 3 to 4. More interesting in this context, perhaps, is the indication that two earlier sections of the work are also atypical of Paine's style, and indeed similar to Lafayette's. To investigate this further, we applied a second function supplied with the stylo package, namely rolling.classify(). For this we again used a block size of 1000 words with a step size of 500. Other settings were system defaults. We used the same comparison texts, except that Hamilton's was omitted, to give a simple 2-class classification problem. The program produced the graph shown here as Figure 12.

Here the portions ascribed to Paine are coloured green, while those ascribed to Lafayette are coloured red. The three horizontal bars of colour indicate the system's first, second and third choice category. Again we see a work mostly attributed to Paine, but with three places where the text resembles writing by Lafayette. The dotted vertical lines labelled 'Q', 'D1' and 'dx' show, respectively, the beginning of Clark's questioned portion; the end of that portion; and the end of the *Declaration* itself. Within these limits, the program finds Lafayette as a plausible author of a majority of the text.

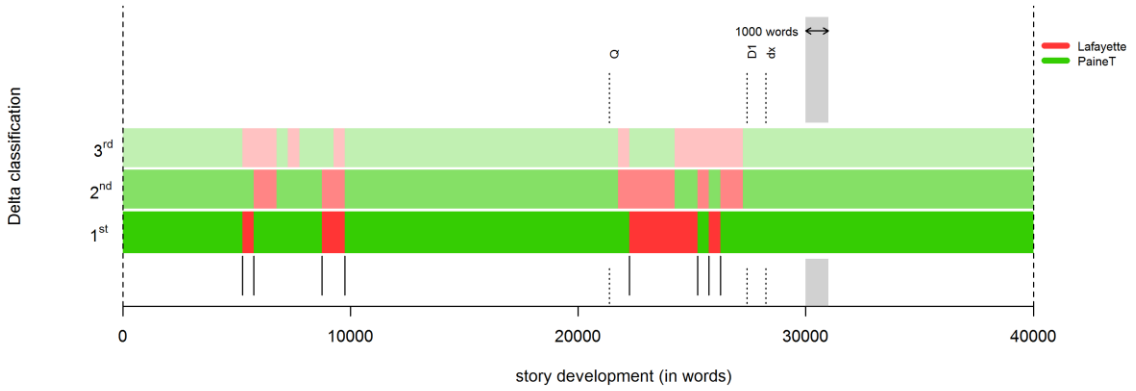

Figure 12. Rolling.classify() using Delta, applied to TROM.

The earlier portions, just after position 5000 and just before position 10000 are also intriguing. To investigate these in particular, we developed a sequential version of the TOCCATA software described earlier which allows us to use TOCCATA's methods in a "rolling" fashion on a single queried text. [See Endnote 2.] This made it possible to supplement the results obtained from stylo with a method that differs significantly from Delta by using a larger feature set and a different way of calculating similarity (the Vote method, described in Appendix 5). It also made it slightly easier to identify the precise locations of the passages rated atypical of Paine.

This Python3 program (slabsim.py) operates in two distinct phases. Phase 1 uses the leave-1-out method of cross-validation in which one text at a time is removed from the training sample and a block of the requested size is taken at random from this left-out text (or two blocks if the text is large enough), to be classified by a model formed from the remaining training texts. Since TOCCATA models produce similarity scores, this permits collection of a series of similarity scores of text blocks not used to form the model when compared with that model. This set is sorted and can be used in the subsequent phase to yield empirical "centile" scores for each block in the queried text.

For example, let us consider a case using deltoid similarity (1/Delta, since TOCCATA works with similarities not dissimilarities). If the training texts yield 80 blocks of the requisite size having deltoid similarity scores ranging from 0.8 to 2.5 with reference to the models built by excluding each of those 80 blocks in turn, these 80 numbers can be sorted and saved for phase 2. Let us further suppose that the 8th and 9th largest of these 80 similarity values are 1.7 and 1.6.

In phase 2, the program runs sequentially through the blocks of the test text. For each block a similarity score is computed comparing the text block to the training-set model. This score is not used directly; instead it is expressed as a centile or percentage with reference to the set generated during the leave-1-out phase. To continue the above example, if the deltoid similarity for a given block were 1.65 this would place it between the 8th and 9th highest of the values retained from phase 1. This beats all but 8 (=72) of the 80 scores, so would be given a centile score of 72/80, or 90 in percentage terms. Thus the centile score for each block in the test text is the percentage of similarity scores obtained during the cross-validation phase which are beaten by, i.e. lower than, the score obtained for the block concerned by applying the classification model to the text under consideration.

A reason for this 2-step procedure is that many different similarity measures are available in TOCCATA, some of which have no natural interpretation, and this puts them all onto a common

scale. Another point is that the initial cross-validation phase gives an unbiased empirical estimate of the expected distribution of similarity scores for unseen blocks of this size, of known category, when compared with a model formed from a training set of that category.

Figure 13 shows results of applying the Vote method to successive 1000-word blocks of TROM, again with a 500-word step size, using 86 training texts, 24 by Paine and 62 by our other authors, including 17 by Lafayette, all categorized for this exercise as "NonPaine". The y-axis, labelled "cent" is a centile score computed as described above. The point to re-emphasize is that these scores are not themselves similarities, but rather indices of how the calculated similarities compare with the similarities computed during the leave-1-out phase when comparing unseen blocks by Paine with the rest of his works.

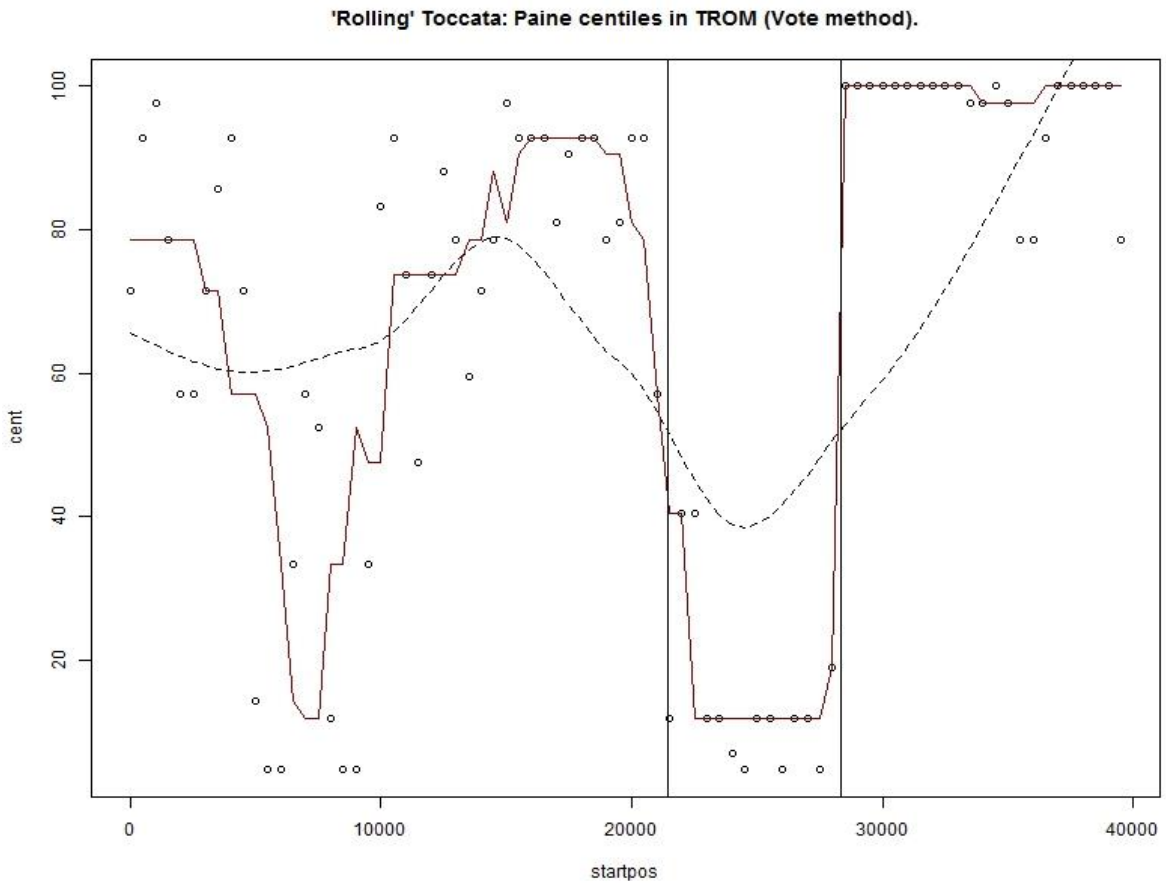

Figure 13. Centile similarity scores of 1000-word blocks of TROM, Vote method.

Here the vertical lines mark the start of the section queried by Clark and the end of the *Declaration*. The heights of the points are centiles assessing how similar each block is to the Paine training-text model, with reference to the values computed during the cross-validation phase. The dashed line results from applying the loess() function in R (R Core Team, 2016) with degree=2 and other parameters as default. This function produces a "smoothed" curve from a series of points, assumed to be spaced at equal intervals. The jagged red line is a different kind of "smooth" curve produced by taking a rolling median along the series, in this case the median of nine items -- the point concerned and the four points on either side of it.

The graph broadly agrees with the results of "rolling" stylo, but we can extract a little more information from it in this case. The queried section appears very clearly as atypical of Paine. Of the

total of 80 points, eighteen fall below the 20th centile, of these, twelve fall within the section demarcated by vertical lines. The rolling median dips dramatically in this section, once again tending to support Clark's contention. Nearly as dramatic is the dip before word-position 10000. This is due to two groups of 3 points which on this measure are unlike Paine's style. The first of these defines a segment of 2000 words starting at position 5000. The second of defines a segment of 2000 words beginning at position 8000.

It is natural to wonder what these sections of the work contain. Referring to the two earlier, smaller, atypical sections as Block 1 and Block 2, and the main atypical section as Block 3, Appendix 7 shows the paragraphs nearest the beginnings and ends of these blocks. All three blocks describe historical events leading up the *Declaration* of Rights: the storming of the Bastille; the expedition to Versailles followed by the journey of King and Queen to Paris; the deliberations of the assemblies, including the historic tennis-court oath. Block 3 ends just prior to a brief quotation -- introduced as an "energetic apostrophe by M. de la Fayette".

To illustrate what sort of plot would result from applying this technique, with the same training set and the same block size and step size, to a different text, we created an artificial composite text-file consisting of the following four files: Lafayette_In America_2b, Queried_Federalist_19 (jointly written by Hamilton & Madison), PaineT_AgeReason_2iii, and Lafayette_toWashington_17790613. None of these were present in the training sample used of 86 texts. The part by Paine of this mixed-author text begins at position 5306 and ends at position 10080.

Of course, concatenating slabs by different writers is not how co-authorship operates in reality, except perhaps in extreme cases. Among other things, the topic focus of each part of this synthetic 'document' is different. Nevertheless, it does give a point of comparison for interpreting the plot produced on a real-life putative mixed-author text such as TROM. Figure 14 shows the result.

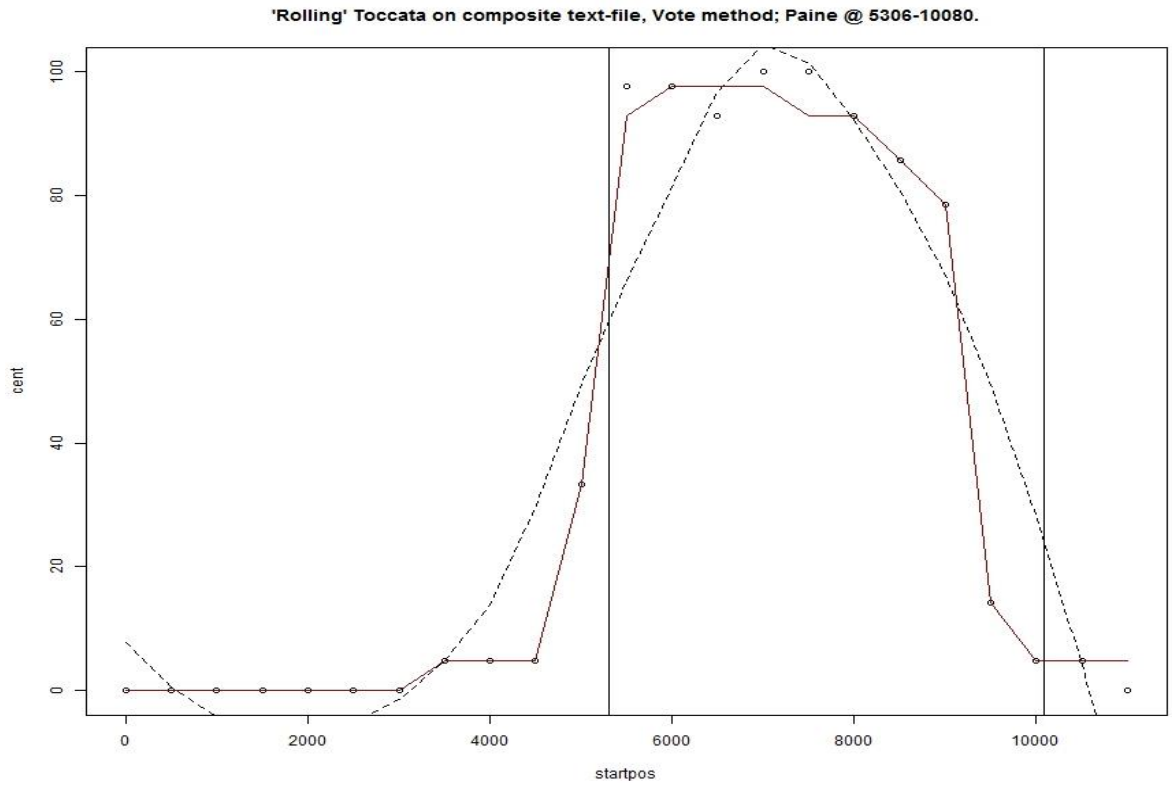

Figure 14. Centile similarity scores to Paine, 1000-word blocks of composite mixed-author text.

As might be expected, this gives a more clear-cut picture. The vertical lines delimit the portion by Paine. Note that the point preceding the left vertical line, at startpos 5000, does contain some text from Tom Paine's section. Likewise, the two points preceding the right vertical line contain text by Lafayette.

What this highlights, in comparison with Figure 13, is that the section of TROM following the *Declaration* resembles the pure-Paine portion of this composite text much more than does the part preceding Clark's questioned passage. It would seem that the latter part of TROM is consistently more similar to Paine's normal style of writing than the part preceding the questioned passage. The earlier part, although it is confidently ascribed to Paine when taken as a whole (subsection 4.2), exhibits signs of mixed authorship throughout.

A plausible explanation for this pattern would be that Lafayette supplied Paine with a historical account of events leading up to the *Declaration*, portions of which were inserted piecemeal &/or heavily edited into the first half of TROM, and almost verbatim into the passage immediately prior to the *Declaration*. The later portion, representing Paine's reflections on the *Declaration* and its historical antecedents, was then written by Paine alone.

This conjecture is compatible with the stylometric evidence, and indeed suggested by that evidence, though it would require corroboration by information external to the texts here examined to be accepted as historical fact.

6. Discussion

The results from the foregoing investigations lead us to believe that the contention of Jonathan Clark (2015) is highly credible. A long passage of text within TROM leading up to the Declaration of Rights itself is very uncharacteristic of Tom Paine's famously distinctive writing style. We have also identified two earlier, shorter but still substantial, passages that are atypical in language with respect to Paine's other writings.

All three passages refer to momentous historical events, so the possibility remains that Paine adapted his habitual language when writing in a genre that he seldom otherwise used. Nevertheless, Paine's familiarity with Lafayette, who had first-hand knowledge of these events, along with the fact that the main queried passage yields results very like those of our known co-authorship (subsection 4.1), leads us to believe that some degree of co-authorship is the most likely explanation. The fact that all three passages assign a central role to Lafayette merely strengthens this belief.

A likely scenario is that Lafayette supplied written descriptions of the events about which he knew more than Paine, which Paine then inserted into his book at points he considered suited to the thrust of his argument, with greater or lesser degrees of editing. Stylometry alone cannot prove this. The putative co-authors are dead. But the stylometric evidence presented here is entirely consistent with such a supposition.

From a methodological viewpoint, we believe this study shows that the TOCCATA package warrants consideration as an addition to the toolkit of stylometric researchers. Furthermore, using "rolling" stylometry, as developed by Eder (2016) in particular, on this problem has confirmed the value of this method within the field of computational stylistics. We consider that our use of more than one "smoothing" method in visualizing the results produced by such procedures represents a contribution to this subfield, and intend to work to further refine such techniques in future, and make the software implementation freely available.

Acknowledgements

We wish to thank Juliana Hessel of The College of New Jersey for collecting some of the electronic texts used in this investigation and for performing a pilot study on this question. We would also like to thank two anonymous reviewers for helpful comments on a previous draft of this paper.

Endnotes

1. In principle the fact that the TOCCATA methods aren't immune to the effects of content and/or register could be ameliorated by weighting the more content-free methods, such as Delta, more heavily when the system is used for authorship attribution; though this was not attempted in the present investigation, as it would require nontrivial amendments to the software.
2. This software, slabsim.py, has not yet been integrated with the TOCCATA package online, since it is presently undocumented and still being refined.

Sources of Software Employed

formulib software:
http://www.richardsandesforsyth.net/software.html
GoTagger:
http://web4u.setsunan.ac.jp/Website/TreeOnline.htm
Python3:
https://www.python.org/downloads/
R:
https://www.r-project.org/
stylo:
https://sites.google.com/site/computationalstylistics/stylo
toccata program:
http://www.richardsandesforsyth.net/software.html

References

Berton G., Petrovic S., Schiaffino R. & Ivanov L. (2016). Examining the Thomas Paine Corpus: Automated Computer Author-Attribution Methodology Applied to Thomas Paine's Writings. In Cleary, S. & Stabell, I.L. (eds.) *New Directions in Thomas Paine Studies*, Basingstoke: Palgrave Macmillan.

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

Burrows, J.F. (1992). Not unless you ask nicely: the interpretive nexus between analysis and information. *Literary and Linguistic Computing*, 7(2): 91-109.

Burrows, J.F. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary & Linguistic Computing*, 17(3), 267-287.

Burrows, J.F. (2010). Never say always again: reflections on the number game. In McCarty, W. (ed.) *Text and Genre in Reconstruction*. OpenBook publishers, 13-35.

Clark, J. (2015). Monuments to liberty. *The Times Literary Supplement*, 16 Sept 2015.

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

Craig H. & Burrows J.F (2012). A collaboration about a collaboration: The authorship of King Henry VI, Part Three. *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*, Ashgate, Farnham, Surrey, 27-65.

Eder, M. (2015). Does size matter? Authorship attribution, small samples, big problem. *Literary & Linguistic Computing*, 30(2), 167-182.

Eder, M. (2016). Rolling stylometry. *Digital Scholarship in the Humanities*, 31(3), 457-469.

Eder, M. (2017). Short samples in authorship attribution: A new approach. In *Digital Humanities 2017: Conference Abstracts* (p. 221–224). Montreal: McGill University. Downloaded from https://dh2017.adho.org/abstracts/341/341.pdf

Eder, M., Rybicki, J. and Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. *R Journal*, 8(1), 107-121.

Forsyth, R.S. (1995). *Stylistic Structures: a Computational Approach to Text Classification*. Unpublished Ph.D. Thesis, University of Nottingham.

Forsyth, R. & Grabowski, Ł. (2015). Is there a formula for formulaic language?. *Poznan Studies in Contemporary Linguistics*, 51(4), 511-549.

Forsyth, R.S. & Holmes, D.I. (1996). Feature-Finding for Text Classification. *Literary and Linguistic Computing*, 11(4): 163-174.

Forsyth, R.S., Holmes, D.I. and Tse, E.K. (1999). Cicero, Sigonio and Burrows: investigating the authenticity of the 'Consolatio'. *Literary and Linguistic Computing*, 14(3): 375-400.

Forsyth, R.S. & Lam, P.W.Y. (2014). Found in translation: To what extent is authorial discriminability preserved by translators? *Literary & Linguistic Computing*, 29(2), 199-217.

Fuller, S. and O'Sullivan, J. (2017). Structure over Style: Collaborative Authorship and the Revival of Literary Capitalism. *Digital Humanities Quarterly*, 11(1). http://www.digitalhumanities.org/dhq/vol/11/1/000286/000286.html

Harris, Z. (1954). Distributional Structure. *Word*, 10:2-3, 146-162.

Hitchens, C. (2006). *Thomas Paine's 'Rights of Man'*. Atlantic Books.

Holmes, D.I. & Crofts, D.W. (2010). The diary of a public man: a case study in traditional and non-traditional authorship attribution. *Literary & Linguistic Computing*, 25(2), 179-197.

Holmes, D.I. & Forsyth, R.S. (1995). The 'Federalist' revisited: new directions in authorship attribution. *Literary & Linguistic Computing*, 10(2), 111-127.

Holmes, D.I., Gordon, L.J. and Wilson, C. (2001). A Widow and her Soldier: Stylometry and the American Civil War. *Literary and Linguistic Computing,* 16(4): 403-420.

Holmes, D.I. & Johnson, E.D. (2012). A stylometric foray into the Anglo-Zulu war of 1879. *English Studies*, 93(3), 310-323.

Holmes, D.I., Robertson, M. and Paez, R. (2001). Stephen Crane and the 'New-York Tribune': A Case Study in Traditional and Non-Traditional Authorship Attribution. *Computers and the Humanities*, 35(3): 315-331.

Hoover, D.L. (2004). Testing Burrows' Delta. *Literary and Linguistic Computing,* 19(4): 453-475.

Jockers, M.L., Witten, D.M. and Criddle, C.S. (2008). Reassessing authorship of the 'Book of Mormon' using Delta and nearest shrunken centroid classification. *Literary and Linguistic Computing,* 23(4): 465-491.

Mosteller, F. & Wallace, D.L. (1984). *Applied Bayesian and Classical Inference: the Case of the Federalist Papers.* New York: Springer. [First edition 1964.]

Oakes, M.P. (2014). *Literary Detective Work on the Computer*. Amsterdam: John Benjamins.

Paine, T. (1791). *Rights of Man*. New York: Dover Publications. [1999 edition.]

Philp, M. (Ed). (2008). *Thomas Paine: Rights of Man, Common Sense and other political writings*. Oxford: Oxford University Press.

R Core Team (2016). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/

Rudman, J. (2016). Some Problems in the Non-Traditional Authorship Attribution Studies of the Dramatic Canon of William Shakespeare: Are they Insurmountable? In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, 662-663.

Rybicki, J. (2012). The great mystery of the (almost) invisible translator. In: Oakes, M.P. & Ji, M. (eds.) Quantitative Methods in Corpus-Based Translation Studies: A practical guide to descriptive translation research, *Studies in Corpus Linguistics 51*. 231–248.

Rybicki, J., Hoover, D. & Kestemont, M. (2014). Collaborative authorship: Conrad, Ford and rolling Delta. *Literary & Linguistic Computing*, 29(1), 422-431.

Sigelman, L. Martindale, C. and McKenzie, D. (1997). The common style of Common Sense. C*omputers and the Humanities*, 30(5), 373-379.

van Dalen-Oskam, K. & van Zundert, J.J. (2007). Delta for Middle Dutch: author and copyist distinction in 'Walewein'. *Literary & Linguistic Computing*, 22(3), 345-362.

Vickers, B. (2004). *Shakespeare, Co-Author*. Oxford: Oxford University Press.

Appendix 1 [Paine Corpus Catalogue]

A zipped file containing these texts is available for research purposes at: http://www.richardsandesforsyth.net/pubs.html.

filename	author	work	source
FranklinB_WesternColonies	Franklin	Plan to settle 2 western colonies	http://www.gutenberg.org
HamiltonA_Continentalist_5	Hamilton	Continentalist No. 5	http://press-pubs.uchicago.edu/founders/
HamiltonA_Continentalist_6	Hamilton	Continentalist No. 6	oll.libertyfund.org
HamiltonA_Federalist_13	Hamilton	Federalist paper 13	http://avalon.law.yale.edu/
HamiltonA_Federalist_25	Hamilton	Federalist paper 25	http://avalon.law.yale.edu/
HamiltonA_Federalist_69	Hamilton	Federalist paper 69	http://avalon.law.yale.edu/
HamiltonA_Federalist_77	Hamilton	Federalist paper 77	http://avalon.law.yale.edu/
HamiltonA_Manufactures_2	Hamilton	Report on Manufactures, part II	http://www.constitution.org/
HamiltonA_MilitarySupplies	Hamilton	Plan for Military Supplies	oll.libertyfund.org
HamiltonA_Pacificus_1	Hamilton	Pacificus No. 1	http://press-pubs.uchicago.edu/founders/
HamiltonA_PlanGovernment	Hamilton	Plan of Government	http://avalon.law.yale.edu
HamiltonA_PublicCredit_1	Hamilton	Public Credit Report 1	www.schillerinstitute.org
HamiltonA_PublicCredit_2	Hamilton	Public Credit Report 2	founders.archives.gov
HamiltonA_PublicLands	Hamilton	Report on Disposition of Lands	oll.libertyfund.org
HamiltonA_toHolt_1778	Hamilton	Letters to Mr Holt, NY Journal	oll.libertyfund.org
HamiltonA_toJay_17790314	Hamilton	Letter to John Jay	http://press-pubs.uchicago.edu/founders/
HamiltonA_toLafayette_ 17891006	Hamilton	to Lafayette, 061089	founders.archives.gov
HamiltonA_toMorris_17820813	Hamilton	Letter to Robert Morris	oll.libertyfund.org
HamiltonA_toWashington_ 17890505	Hamilton	Letter to General Washington	http://press-pubs.uchicago.edu/founders/
HamiltonA_VindicationCongress	Hamilton	Full Vindication of Measures	founders.archives.gov
JayJ_Federalist_4	Jay	Federalist paper 4	http://avalon.law.yale.edu/
JayJ_Federalist_64	Jay	Federalist paper 64	http://avalon.law.yale.edu/
JeffersonT_Autobiography_1	Jefferson	Autobiography, part 1	en.wikisource.org
JeffersonT_Inaugural_1805	Jefferson	Second Inaugural address	en.wikisource.org

JeffersonT_Independence	Jefferson	Declaration of Independence	www.archives.gov
JeffersonT_NotesVirginia	Jefferson	Notes on State of Virginia, 17-20	oll.libertyfund.org
JeffersonT_StateofUnion_1801	Jefferson	First State of Union address	en.wikisource.org
JeffersonT_StateofUnion_1808	Jefferson	Eighth State of Union address	en.wikisource.org
JeffersonT_toGovernorDunmore	Jefferson	Address to Governor Dunmore	oll.libertyfund.org
JeffersonT_toHenry_17790327	Jefferson	Letter to Governor Patrick Henry	oll.libertyfund.org
JeffersonT_toMadison_ 17851028	Jefferson	Letter to Madison	en.wikisource.org
JeffersonT_toMason_17910204	Jefferson	Letter to Gerorge Mason	oll.libertyfund.org
JeffersonT_TonnageLaw	Jefferson	Report on Tonnage Law	oll.libertyfund.org
JeffersonT_UsefulArts	Jefferson	Draft Bill to Promote Useful Arts	oll.libertyfund.org
MadisonJ_BillofRights_1789	Madison	Bill of Rights	founders.archives.gov
MadisonJ_Federalist_42	Madison	Federalist paper 42	http://avalon.law.yale.edu/
MadisonJ_Federalist_48	Madison	Federalist paper 48	http://avalon.law.yale.edu/
MadisonJ_Inaugural_1809	Madison	Inaugural Address 1809	http://www.presidency.ucsb.edu/
MadisonJ_Inaugural_1813	Madison	Inaugural Address 1813	http://www.presidency.ucsb.edu/
MadisonJ_Money	Madison	Observations on Money	founders.archives.gov
MadisonJ_StateofUnion_1809	Madison	State of Union address 1809	http://www.gutenberg.org
MadisonJ_StateofUnion_1816	Madison	State of Union address 1816	http://www.gutenberg.org
MadisonJ_toJefferson_ 17871024	Madison	Letter to Jefferson	http://oll.libertyfund.org/
MadisonJ_toWashington_ 17870930	Madison	Letter to General Washington	http://oll.libertyfund.org/
MadisonJ_UniversalPeace	Madison	Universal Peace in Nat. Gazette	founders.archives.gov
Lafayette_InAmerica_1a	Lafayette	First Voyage and Campaign	http://www.gutenberg.org
Lafayette_InAmerica_1b	Lafayette	First Voyage and Campaign	http://www.gutenberg.org
Lafayette_InAmerica_1c	Lafayette	First Voyage and Campaign	http://www.gutenberg.org
Lafayette_InAmerica_1d	Lafayette	First Voyage and Campaign	http://www.gutenberg.org
Lafayette_InAmerica_1e	Lafayette	First Voyage and Campaign	http://www.gutenberg.org

Lafayette_InAmerica_1f	Lafayette	First Voyage and Campaign	http://www.gutenberg.org
Lafayette_InAmerica_2a	Lafayette	Second Voyage and Campaign	http://www.gutenberg.org
Lafayette_InAmerica_2b	Lafayette	Second Voyage and Campaign	http://www.gutenberg.org
Lafayette_MapExplanation1781	Lafayette	Appendix Explaining Campaign Map	http://www.gutenberg.org
Lafayette_toDayen_17780911	Lafayette	Letter to Duc D'Ayen, wife's father	http://www.gutenberg.org
Lafayette_toRochTern_ 17800809	Lafayette	Letter to Rochambeau & Ternay	http://www.gutenberg.org
Lafayette_toVergennes_ 17790718	Lafayette	Letter to Comte de Vergennes	http://www.gutenberg.org
Lafayette_toVergennes_ 177908	Lafayette	Letter to Comte de Vergennes	http://www.gutenberg.org
Lafayette_toVergennes_ 17800202	Lafayette	Letter to Comte de Vergennes	http://www.gutenberg.org
Lafayette_toWashington_ 17790111	Lafayette	Letter to General Washington	http://www.gutenberg.org
Lafayette_toWashington_ 17790613	Lafayette	Letter to General Washington	http://www.gutenberg.org
Lafayette_toWashington_ 17791007	Lafayette	Letter to General Washington	http://www.gutenberg.org
Lafayette_toWashington_ 17801030	Lafayette	Letter to General Washington	http://www.gutenberg.org
Lafayette_toWife_17780913	Lafayette	Letter to Marie Adrienne, wife	http://www.gutenberg.org
JohnsonS_Idler103_1760	Johnson	Last Idler Essay by Dr Johnson	http://www.gutenberg.org
LincolnA_Gettysburg1863	Lincoln	Lincoln's Gettysburg Address	http://avalon.law.yale.edu/
Queried_AfricanSlavery	Queried	African Slavery	http://thomaspaine.org
Queried_Federalist_19	Both	Federalist paper 19	http://avalon.law.yale.edu/
Queried_Federalist_20	Both	Federalist paper 20	http://avalon.law.yale.edu/
Queried_Federalist_52	Queried	Federalist paper 52	http://avalon.law.yale.edu/
Queried_Federalist_57	Queried	Federalist paper 57	http://avalon.law.yale.edu/
Queried_UnhappyMarriages	Queried	Reflections on Unhappy Marriages	http://thomaspaine.org
Shax_sonn109	Queried	Shakespeare's sonnet 109	http://www.gutenberg.org
Theo_749f	Queried	Theo Van Gogh to brother in French	http://vangoghletters.org/vg/letters/
TROM_Declaration	Anon	TROM Declaration of Rights	http://www.gutenberg.org

TROM_Mid1	Queried	TROM Queried Part1	http://www.gutenberg.org
TROM_Mid2	Queried	TROM Queried Part2	http://www.gutenberg.org
PaineT_AbbeRaynal_1	Paine	Letter to Abbe Raynal 1	http://thomaspaine.org
PaineT_AbbeRaynal_2	Paine	Letter to Abbe Raynal 2	http://thomaspaine.org
PaineT_AgeReason_1vii	Paine	Age of Reason 1vii	http://www.gutenberg.org
PaineT_AgeReason_1xvii	Paine	Age of Reason 1xvii	http://www.gutenberg.org
PaineT_AgeReason_2iii	Paine	Age of Reason 2iii	http://www.gutenberg.org
PaineT_AgrarianJustice_1	Paine	Agrarian Justice 1	http://thomaspaine.org
PaineT_AgrarianJustice_2	Paine	Agrarian Justice 2	http://thomaspaine.org
PaineT_AmericanCrisis_1	Paine	American Crisis 1	http://thomaspaine.org
PaineT_AmericanCrisis_13	Paine	American Crisis 13	http://thomaspaine.org
PaineT_AmericanCrisis_2	Paine	American Crisis 2	http://thomaspaine.org
PaineT_AmericanCrisis_7	Paine	American Crisis 7	http://thomaspaine.org
PaineT_AmericanCrisis_Sup1	Paine	American Crisis Supernumery 1	http://thomaspaine.org
PaineT_CommonSense_1	Paine	Common Sense 1	http://thomaspaine.org
PaineT_CommonSense_4	Paine	Common Sense 4	http://thomaspaine.org
PaineT_Fisheries_3	Paine	Newfoundland Fisheries, part 3	www.thomaspaine.org
PaineT_Government	Paine	Dissertations on Government	www.thomaspaine.org
PaineT_PublicGood_A	Paine	Public Good, up to Diagram 2	www.bartleby.com
PaineT_QuakerEpistle	Paine	Epistle to Quakers	www.thomaspaine.org
PaineT_Rights_2i	Paine	TROM Part 2 Intro & Chapter 1	http://www.gutenberg.org
PaineT_Rights_2iv	Paine	TROM Part 2 Chapter 4	http://www.thomaspaine.org/
PaineT_Rights_Conclusion	Paine	TROM Conclusion	http://www.gutenberg.org
PaineT_Rights_Miscellaneous	Paine	TROM Miscellaneous Chapter	http://www.gutenberg.org
PaineT_Rights_Observations	Paine	TROM Observations	http://www.gutenberg.org
PaineT_Rubicon	Paine	Prospects on the Rubicon	www.thomaspaine.org
PaineT_toFranklin_17780516	Paine	Letter to Benjamin Franklin	http://www.thomaspaine.org/
PaineT_toJefferson_18050125	Paine	Letter to Jefferson	www.thomaspaine.org
PaineT_toMonroe_17941013	Paine	Letter to James Monroe	www.thomaspaine.org
TROM_Preface	Paine	Preface to Rights of Man, Part I	http://www.gutenberg.org
TROM_Rights_1	Paine	Up to: The despotism of Louis XIV	http://www.gutenberg.org

These texts were only lightly pre-processed before analysis. Normalization was limited to standardizing the character-representations of apostrophes, dashes, hyphens and quotation marks. Embedded quotations were not deleted except in the single case of *TROM_Declaration* itself, which we believe is the longest quoted passage in any of these works. This was extracted into an individual file (listed above).

Note: Six of the 10 letters by Marquis de Lafayette, i.e. all except those to General Washington, were written originally in French. They were translated into English for the 1837 edition of Lafayette's works, but the editor does not state by whom. Thus it is safe to treat them as not being by Thomas Paine, although treating them as definitely by Lafayette is problematic. In this connection, however, it is worth noting the fact that both Rybicki (2012) and Forsyth & Lam (2014) have found that attributing authorship in translation is, surprisingly, almost as easy as in the original language.

Appendix 2 [Reflexive Co-authorship Corpus Catalogue]

filename	doctype	authors	textname
CO_Feds_1995.txt	CO	Holmes & Forsyth	Federalist Revisited
CO_Formulang_2015.txt	CO	Forsyth & Grabowski	Formulaic Language
CO_FoundTran_2014	CO	Forsyth & Lam	Found in Translation (van Goghs)
CO_PublicMan_2010.txt	CO	Holmes & Crofts	Diary of a Public Man
CO_Zulu_2012.txt	CO	Holmes & Johnson	Anglo-Zulu War
DH_Authorship_1993.txt	DH	Holmes	Authorship Attribution
DH_Crane_2001.txt	DH	Holmes	Stephen Crane study
DH_EvoStylo_1998.txt	DH	Holmes	Evolution of Styometry
DH_LettAmerica_2008.txt	DH	Holmes	Letter from America
DH_LitStyle_1985.txt	DH	Holmes	Analysis of Literary Style
DH_Mormons_1992.txt	DH	Holmes	Mormon Scriptures
DH_Oxford_1992.txt	DH	Holmes	Oxford Conference Report
DH_Pickett_2012.txt	DH	Holmes	Pickett Letters
DH_ProphetVoice_1991.txt	DH	Holmes	Vocabulary Richness & Prophetic Voice
DH_Sorbonne_1994.txt	DH	Holmes	Sorbonne Conference Report
DH_Stylometry_2003.txt	DH	Holmes	Introduction to Stylometry
DH_Widow_2002.txt	DH	Holmes	Widow & Soldier
DH_Xmas_2004.txt	DH	Holmes	Christmas Message
DH_Zululand_2009.txt	DH	Holmes	Zululand Field Trip
MO_Authorid_2014.txt	MO	Oakes	Author Identification
RF_Anns_1993.txt	RF	Forsyth	Artificial Neural Nets
RF_Beaghelp_1988.txt	RF	Forsyth	PC/Beagle Help File
RF_BullBear_1989.txt	RF	Forsyth	Bulls & Bears & Floppy Discs, FT
RF_Cons1to4_1999.txt	RF	Forsyth	Consolatio Paper, parts 1 to 4
RF_DDDD_1997.txt	RF	Forsyth	Deriving Document Descriptors
RF_Exclusion_2003.txt	RF	Forsyth	Authorship Exclusion

RF_Experts_1983.txt	RF	Forsyth	Expert System Architecture
RF_Foreword_1994.txt	RF	Forsyth	Foreward to Chorafas book
RF_Ioga_1996.txt	RF	Forsyth	Instance-Oriented Genetic Algorithm
RF_MLchapter1_1988.txt	RF	Forsyth	Machine Learning, chapter 1
RF_NeuroNottm_1990.txt	RF	Forsyth	Neural Learning Trials, Nottm report
RF_Ockhams_1992.txt	RF	Forsyth	Ockham's Razor
RF_Phd7_1995.txt	RF	Forsyth	Thesis Chapter 7
RF_PopFlops_2000.txt	RF	Forsyth	Pops & Flops
RF_Robayes_1996.txt	RF	Forsyth	Robust Bayesian Classifier
RF_Stylochron_1999.txt	RF	Forsyth	Stylochronometry with substrings
RF_Teaboat_2012.txt	RF	Forsyth	Teaboat Software Overview
RF_TeskeyRev_1983.txt	RF	Forsyth	Review of Teskey Text Processing
RF_TextCat_2008.txt	RF	Forsyth	Authorship & Text Classification
RF_Wism_1991.txt	RF	Forsyth	Grounded Morality
DH_Fed2_1995.txt	DH	Holmes	Feds Part 2 (& a little of Part 1)
DH_Fed3_1995.txt	DH	Holmes	Feds Part 3
DH_Fed5_1995.txt	DH	Holmes	Feds Conclusion
RF_Fed1_1995.txt	RF	Forsyth	Feds Part 1 (most of)
RF_Fed4_1995.txt	RF	Forsyth	Feds Part 4

These texts were pre-processed in the same way as those listed in Appendix 1; i.e. apostrophes, dashes, hyphens and quotation marks were standardized. In addition, in those papers where Reference Lists were present, they were removed.

Appendix 3 [Words used in Burrows-style analysis of Paine & contemporaries]

Retained words (n=100)

the	of	to	and	a	in	that	it	is	be
as	which	for	by	not	with	have	this	or	on
they	are	but	was	from	their	will	at	an	would
all	them	has	had	been	may	if	one	other	were
its	any	no	more	government	than	who	there	so	those
can	what	such	general	should	shall	some	time	every	upon
when	into	state	only	power	people	same	these	must	made
part	great	country	could	being	without	new	two	now	under
first	most	might	own	out	do	then	make	against	each
much	very	before	well	necessary	place	present	good	too	either

Excluded words (n=44)

i	he	his	we	you	our	my	your	him	man
states	right	her	us	war	nation	me	america	public	men
france	england	bank	whole	constitution	itself	therefore	money	mr	case
up	interest	congress	national	because	cannot	themselves	de	she	ought
act	rights	united	between						

Appendix 4 [Word used in Burrows-style analysis of present co-authors]

Retained words (n=100)

the	of	and	in	to	a	is	that	as	for
this	by	it	be	on	are	with	from	which	not
have	at	but	one	an	than	has	all	they	most
also	so	being	was	or	two	these	each	more	i
can	such	other	their	only	well	very	been	first	into
would	some	had	their	its	them	about	were	used	between
number	if	both	work	will	then	when	out	should	time
known	what	we	may	found	who	use	any	many	over
data	using	1	no	different	written	given	however	could	how
words	analysis	three	same	problem	where	here	text	2	test

Excluded words (n=44)

his	e	word	he	set	new	function	case	present	our
results	approach	features	study	method	table	author	based	example	texts
n	sample	program	authors	variables	thus	length	authorship	corpus	style
samples	textual	language	frequency	p	letters	book	papers	vocabulary	training
principal	rules	poems	diary						

Appendix 5 [Text Classification Methods used by TOCCATA]

In the studies reported above, the setting of parameter "wordonly" was 1, meaning that only tokens beginning with an alphanumeric character were considered. Thus in the following "words" and "tokens" are interchangeable (provided that numbers are considered as words). In particular, punctuation symbols were ignored, although in many circumstances -- e.g. when known to be under authorial control -- they can be effective stylometric markers.

Deltoid

Module Deltoid is an implementation of Burrows's delta (Burrows, 2002) which has become a standard technique in authorship attribution studies. In a nutshell, this method first finds the most frequent N word tokens in the corpus; then computes the means and standard deviations of the relative usage rates of these words across the various documents of the corpus. This allows it to consider the mean usage rates of these words in each category as a model of that class. To compare a single text with a class model, it computes the absolute z-scores of all these words and averages them, a z-score being computed by subtracting the usage rate of the word under consideration in the text from the mean rate in the class model and dividing this difference (ignoring sign) by the standard deviation of that word in the corpus as a whole. This process yields a mean absolute z-score, which is a dissimilarity measure. Because TOCCATA works with similarities, these mean dissimilarities (d_i) are converted to similarities as $1.0/d_i$. The number, N, of most-frequent words to employ is a user-selectable parameter but if this is absent the system sets N to be the square root of the vocabulary size V (i.e. total different vocabulary items, not total running tokens), which is usually a reasonable choice. In the present study the default, N=round(sqrt(V)), was used.

MAWS

This library module implements a method inspired by what Mosteller & Wallace, in their classic work (1964/1984) on the disputed Federalist papers, call their "robust Bayesian analysis"; hence MAWS, Mosteller And Wallace System (Mosteller & Wallace, 1984). The implementation in TOCCATA is a slight revision of the software originally in Forsyth's PhD thesis (Forsyth, 1995) to automate this

approach. Essentially this is a naive Bayesian classifier using frequent word tokens. It takes 2 parameters, toks2get and multivox (default values 144 and 1.618034 respectively) and computes the rounded value of toks2get multiplied by multivox. It then takes the resulting number of the most common words in the corpus, according to document frequency, and reduces them to toks2get again by discarding the least discriminatory items. Then Bayes factors are computed from the training data for each remaining token according to how often the relative frequency of that token exceeds the median frequency of that token in the whole corpus.

Vote

Module Vote is exceptional in that it actually uses every single word-type in the training corpus as a feature. The 'model' developed for classification consists simply of frequency-tables for every text category, containing the frequency of that word in that category. To classify a new text, the frequency of every word in that text is counted; then, for each category, a similarity score is computed by adding a 'vote' by each word in the text to a running total. The vote is either sqrt(f) or 0, where f is the frequency of that word in the text being classified , according to whether the word's relative frequency is more common in the category concerned than in the corpus as a whole, or not. The category yielding the highest total is chosen. This simple procedure was intended merely to establish a baseline, but various trials have shown that it works rather well, perhaps because information from the whole vocabulary structure is utilized.

Tokspans

This method attempts to capture some of the information inherent in *syllexis*, the co-occurrence of words in close proximity, and especially sequential co-occurrence. Being aimed at authorship attribution it starts, conventionally enough, by finding the most common words in the corpus. The number kept can be set by the user but by default is the square root of the vocabulary size, rounded to an integer, with an enforced maximum of 1024. Frequency is defined not by gross count but by the number of snippets in which a word occurs. Default snippet size is 144 tokens. Thus dispersed words that occur throughout the corpus are favoured over "bursty" words that occur often but only in few segments. This preliminary word-selection can be considered step 0 of the overall procedure, after which follow a further four steps: (1) accumulate; (2) eliminate; (3) correlate; (4) discriminate. Steps 1 and 2 constitute the model-building phase; steps 3 and 4 form the model-using phase.

Step 1 goes through the texts examining each segment of S consecutive words, where S is a parameter called spansize, set by the user. With spansize=3, as in the trials reported above, the triplet "by the user", for example, would be examined and -- presuming "by" and "the" were in the frequent word list but "user" was not -- the tuple ("by","the") would be retained with 1 added to its occurrence tally. This typically generates a very large number of token tuples, most of which are shorter than spansize in length. In fact, the zero-length tuple is normally one of the most frequent -- implying a segment of S tokens all absent from the high-frequency list. Most of these tuples are winnowed out in step 2. In this step any such tuples occurring in fewer than 2 texts are removed, and, more significantly, the tuples are rated by how strongly they are associated with each category, with only the most distinctive retained. The default mode of calculating distinctiveness is to compute (r1-r2)*sqrt(df), where r1 is the rate in the category under consideration, r2 is the rate in all other categories combined and df is the number of documents of the category under consideration in which the token occurs. The highest-scoring items in each category are kept. The number to keep for each category is the rounded square root of that category's vocabulary size. The lists for each category are then merged and become a list of distinctive features. In phase 3 this distinctive-feature list is used to calculate a correlation coefficient (rank correlation by default) for each test text with each category in the following manner. The frequencies of every distinctive feature in the text under consideration, including zero, form one vector which is correlated with vectors of the total frequencies of those features in each category, giving C correlations, where C is the number of

categories. In step 4, the highest of those correlations is chosen to assign a category to the text concerned.

Tagsets

This method actually uses the same software library as Tokspans, above, with different parameter settings. Among the chief differences are that sets of tokens were collected, not tuples, i.e. that sequence within spans was ignored, and that part-of-speech tags (as assigned by the GoTagger; see Appendix 6) were used as tokens not orthographic words. In addition, the spansize parameter was set to 5. Hence, again, many sets chosen as features consist of less than 5 items. For example, the quintet "the point of this method" would be tagged as "dt nn of dt nn" and would thus generate the potential feature ['dt', 'nn', 'of'], since sets don't include replications. (On output, the items are listed in alphabetic order, which may not reflect the order in the original text.)

Taverns

Taverns is another method exploiting sequential information. The name stands for (Textual Affinity Values Employing Repeated N-gram Sequences). It employs a technique borrowed from the formulib package, developed to explore formulaic language, which is available at the website below.
http://www.richardsandesforsyth.net/software.html
Essentially this method allows short n-grams to overlap. In the experiments reported above, n-grams of length from 2 to 4 items were generated for each text category and the most frequent at each size retained. Thus, for instance, in Tom Paine's writings the following 4- and 3-grams were kept, among many others:

('of', 'the', 'united', 'states')
('united', 'states', 'of', 'america')
('the', 'people', 'of').

When classifying a fresh text using the list of frequent n-grams in each category, the occurrences of each are not just counted. Rather, the proportion of the text covered jointly by all the n-grams in the list is computed. Thus, if the text contained the 8-gram "the people of the united states of america" that would count as 8 tokens covered by a combination of the three items above: "the people of" would be marked as covered by the 3-gram; "of the united states" would be covered by the first 4-gram; "united states of america" would be covered by the second 4-gram. The fact that some words were covered twice wouldn't matter. The eventual similarity score would only depend on what proportion of tokens in the text being classified had been covered overall, not on how many times each word was covered nor how many times particular n-grams had been found in the text. Note that any word can appear in an n-gram: in our experiments there was no preliminary exclusion of words on the basis of their individual frequencies.

Appendix 6 [GoTagger Part-of-speech Tags, after post-processing]

GoTagger codes (after post-processing) :	
Tagcode	**Description**
BE	Any form of verb "to be"
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition/subord. conjunction

JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
OF	of
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRPS	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund/present participle
VBN	Verb, past participle
VBP	Verb, non-3rd ps. sing.
VBZ	Verb, 3rd ps. sing. Present
WDT	wh-determiner
WP	wh-pronoun
WPS	Possessive wh-pronoun
WRB	wh-adverb
#	Pound sign & miscellaneous punctuation
$	Dollar sign
.	Sentence-final punctuation
,	Comma
:	Colon, semi-colon
(	Left bracket character
)	Right bracket character

The GoTagger software, by Kazuaki Goto, is freely available at the website below:
http://web4u.setsunan.ac.jp/Website/TreeOnline.htm
It is not the most modern part-of-speech software, and its accuracy for computational linguistics is not state-of-the-art; but it is fast, free and consistent, the last being the most important attribute in the context of authorship attribution. We have written a short post-processing program to deal with some of the quirks of GoTagger that caused problems for TOCCATA's tokenization routine, so the

codes above are not quite identical with those listed on the website above. (Punctuation was not used in the experiments reported, so the last 7 symbols above were ignored.)

Appendix 7 [Initial and final paragraphs of three TROM text blocks identified as more like Lafayette than Paine by rolling stylometry.]

Position	Paragraph
Block 1 start	As Mr. Burke has passed over the whole transaction of the Bastille (and his silence is nothing in his favour), and has entertained his readers with reflections on supposed facts distorted into real falsehoods, I will give, since he has not, some account of the circumstances which preceded that transaction. They will serve to show that less mischief could scarcely have accompanied such an event when considered with the treacherous and hostile aggravations of the enemies of the Revolution. [....]
Block 1 end	That the Bastille was attacked with an enthusiasm of heroism, such only as the highest animation of liberty could inspire, and carried in the space of a few hours, is an event which the world is fully possessed of. I am not undertaking the detail of the attack, but bringing into view the conspiracy against the nation which provoked it, and which fell with the Bastille. The prison to which the new ministry were dooming the National Assembly, in addition to its being the high altar and castle of despotism, became the proper object to begin with. This enterprise broke up the new ministry, who began now to fly from the ruin they had prepared for others. The troops of Broglio dispersed, and himself fled also.
Block 2 start	I give to Mr. Burke all his theatrical exaggerations for facts, and I then ask him if they do not establish the certainty of what I here lay down? Admitting them to be true, they show the necessity of the French Revolution, as much as any one thing he could have asserted. These outrages were not the effect of the principles of the Revolution, but of the degraded mind that existed before the Revolution, and which the Revolution is calculated to reform. Place them then to their proper cause, and take the reproach of them to your own side. [....]
Block 2 end	During the latter part of the time in which this confusion was acting, the King and Queen were in public at the balcony, and neither of them concealed for safety's sake, as Mr. Burke insinuates. Matters being thus appeased, and tranquility restored, a general acclamation broke forth of Le Roi a Paris- Le Roi a Paris- The King to Paris. It was the shout of peace, and immediately accepted on the part of the King. By this measure all future projects of trapanning the King to Metz, and setting up the standard of opposition to the constitution, were prevented, and the suspicions extinguished. The King and his family reached Paris in the evening, and were congratulated on their arrival by M. Bailly, the Mayor of Paris, in the name of the citizens. Mr. Burke, who throughout his book confounds things, persons, and principles, as in his remarks on M. Bailly's address, confounded time also. He censures M. Bailly for calling it "un bon jour," a good day. Mr. Burke should have informed himself that this scene took up the space of two days, the day on which it began with every appearance of danger and mischief, and the day on which it terminated without the mischiefs that threatened; and that it is to this peaceful termination that M. Bailly alludes, and to the arrival of the King at Paris. Not less than three hundred thousand persons arranged themselves in the procession from Versailles to Paris, and not an act of molestation was committed during the whole march.

Block 3 start	As we are to view this as the first practical step towards the Revolution, it will be proper to enter into some particulars respecting it. The Assembly of the Notables has in some places been mistaken for the States-General, but was wholly a different body, the States-General being always by election. The persons who composed the Assembly of the Notables were all nominated by the king, and consisted of one hundred and forty members. But as M. Calonne could not depend upon a majority of this Assembly in his favour, he very ingeniously arranged them in such a manner as to make forty-four a majority of one hundred and forty; to effect this he disposed of them into seven separate committees, of twenty members each. Every general question was to be decided, not by a majority of persons, but by a majority of committee, and as eleven votes would make a majority in a committee, and four committees a majority of seven, M. Calonne had good reason to conclude that as forty-four would determine any general question he could not be outvoted. But all his plans deceived him, and in the event became his overthrow. [....]
Block 3 end	Having now traced the progress of the French Revolution through most of its principal stages, from its commencement to the taking of the Bastille, and its establishment by the Declaration of Rights, I will close the subject with the energetic apostrophe of M. de la Fayette- "May this great monument, raised to Liberty, serve as a lesson to the oppressor, and an example to the oppressed!"