

The Authorship of The American Declaration of Independence

Peter W.H. Smith, David A. Rickards

Abstract. Thomas Jefferson, the architect and author of the American Declaration of Independence (ADOI) is revered, and yet, even during his lifetime questions were raised about his authorship [1]. One name linked with the ADOI is Thomas Paine, the author of *Common Sense and The Rights of Man*. This study uses discriminant analysis to and Burrow's Delta scores [2] which reveal that both Jefferson and Paine exhibit a consistent style of writing. A word set is then created to discriminate between Jefferson and Paine using a hybrid genetic algorithm. From this an n-dimensional convex hull is used as test for authorship of the ADOI itself. Further tests are carried out based on sentence fragments. The study takes into account the sources of the ADOI including *The Virginia Constitution* and *The Summary View of The Rights of America* [8]. Results are based on an analysis of all known texts by Thomas Jefferson and Thomas Paine written *prior to* the signing of the ADOI. Test results indicate that Thomas Paine is the possible author of this historic document.

1 INTRODUCTION

Even during his lifetime, Thomas Jefferson was forced to defend his position as author of ADOI [1]. [3] is a very early attempt to use a form of systematic analysis to test the authorship of the ADOI.

At least seven different versions of the ADOI are known to exist [1], although fragments of an even earlier version have been found [4]. These versions are:

1. A Copy in the handwriting of Thomas Jefferson – known as *The Rough Draft*. (Massachusetts Historical Society, Boston).
2. A Copy in the handwriting of John Adams – thought to be an early copy of *The Rough Draft*. (Also at Massachusetts Historical Society, Boston).
3. A Further copy in the handwriting of Thomas Jefferson. (New York Public Library).
4. A Draft in the handwriting of Thomas Jefferson (American Philosophical Society, Philadelphia).
5. The Declaration as printed by Dunlap under the orders of Congress.
6. The Declaration written out in the corrected journal.
7. The Declaration on parchment at the Department of State.

At first glance, the ADOI looks unsuitable for an authorship attribution study, because such an important document would almost certainly have been much changed by Committee and The Continental Congress. Moreover, the evidence for Jefferson's authorship, would on the face of it, appear to be incontrovertible. He was appointed to draft the declaration and two texts known to have been written by Jefferson appear to be sources for substantial sections of the ADOI. However, a copy was produced prior to any amendments by Congress. Even the earliest known version appears to have been copied from an earlier copy now lost [1,4]. Furthermore, during his own lifetime Jefferson felt compelled to comment on his authorship stating that he drafted the ADOI, *without reference to pamphlet or book* - a comment that has puzzled scholars ever since.

The version chosen for this analysis is known as *The Adams' Copy*. There has been much speculation about it, though it seems clear, from a detailed analysis [1], that it is a copy of *The Rough Draft*. That it is in John Adams' handwriting, is not in dispute as his handwriting is confirmed in the biography by his grandson [5]. John Adams was also known to have sent this copy to his wife which also supports this view.

There are differences between *The Adams' Copy* and *The Rough Draft* that are puzzling. They are minor, as they consist largely of punctuation, but there are also a few minor spelling variations and couple of small grammatical changes. It has been suggested by [6] that the differences may be accounted for if *The Adams' copy* was written down by dictation. This simple explanation appears highly plausible, but in no way affects our study. Of the spelling variants between *The Adams' Copy* and *The Rough Draft*, none were found to be used by either Thomas Paine or Thomas Jefferson, but it was discovered that the spelling variant *tryal* was used by John Adams - in Clarendon No. 3 [7] - a small, but possibly significant finding that supports Whissell's theory.

The Adams' Copy was therefore chosen, because it is free of corrections made by Jefferson, Benjamin Franklin or by Congress and it is also possible, within reason, to date the evolution of the document. For the remainder of this paper, unless otherwise stated, we refer to *The Adams' Copy* of the ADOI.

2 METHOD

The study was carried out in four distinct phases:

- A study of the consistency of Jefferson's and Paine's writing using discriminant analysis and for comparison, Burrows' Delta method [2].
- The development of a word vector to separate the two authors, developed using a hybrid genetic algorithm.
- The application of the word vector to the ADOI.

¹ Department of Computing, City University, Northampton Square, London, EC1V 0HB. Email: peters@soi.city.ac.uk

² 18001 Euclid Avenue, Cleveland, OH 44112-1105, USA. Email: docrick@petalk.com

- A comparison using vocabulary, phrases and grammatical constructs.

2.1 The Writing of Thomas Jefferson and Thomas Paine

Thomas Jefferson's writing was extensive and varied. The authoritative source is [8]. For the first part of the study we chose a mixture of letters and documents written by him, in addition to his autobiography. Thomas Paine wrote a number of full-length texts as well as a series of articles. The authoritative source for his texts is [9]. For the first part of this study, we chose three of his full-length works as well as the set of essays known collectively as *The American Crisis*. This gave us seven texts for Jefferson and five by Paine, collectively over 500,000 words.

The Jefferson texts are as follows:

1. Jefferson letters and documents 1760-1786 [j1]
2. Jefferson letters and documents 1786-1792 [j2]
3. Jefferson letters and documents 1792-1803 [j3]
4. Jefferson letters and documents 1803-1811 [j4]
5. Jefferson letters and documents 1812-1817 [j5]
6. Jefferson letters and documents 1817-1822 [j6]
7. Jefferson's autobiography 1821 [auto]

The following works by Paine were chosen:

1. Common Sense (1775) [sense]
2. The American Crisis (1776-1778) [crs1]
3. The American Crisis pt 2 (1778-1783) [crs2]
4. Rights of Man (1792) [rom]
5. The Age of Reason (1795) [aor]

The texts were converted into a standard form for consistency removing all extraneous text, quotes and foreign language sections. Jefferson's letters contain several passages in languages other than English and Paine's *Age of Reason*, for example contains a large number of biblical quotations.

For the purposes of the discriminant analysis, the texts were broken up into blocks of 2500 words. The discriminant analysis used the top 20 function words.

The results were checked by running Burrow's Delta method [2]. For this, we constructed a large text corpus of 2.4 million words, comprising texts by contemporaries of Paine and Jefferson, both British and American. The authors used in this corpus were: Alexander Hamilton, James Madison, Benjamin Franklin, John Adams, Adam Ferguson, Adam Smith, Edmund Burke, Edward Gibbon, Frances Brooke, Gilbert White, Horace Walpole, James Boswell, Robert Kerr, Samuel Adams, Samuel Johnson, Thomas Clarkson and William Beckford. The writing consisted of political speeches and documents, letters and longer texts, all of which are representative of the type of text written by Paine and Jefferson.

The discriminant analysis on the Jefferson texts revealed that texts j1/j2 and j4/5/6 were difficult to separate out. Additionally four of the five Paine texts showed a great deal of consistency of style by word frequency. However, one text for Jefferson and one for Paine appeared slightly problematic using this method – Jefferson's Autobiography and Paine's *Age of Reason*. The reason for the apparent difference of style in Jefferson's autobiography wasn't clear but Jefferson's writing spanned over 50 years and his autobiography was written in

1821. Paine's *Age of Reason* also exhibited different results, but it too was written later in Paine's life³ – a considerable time after the ADOI.

To compute the delta scores, the method described in [2] was followed. The mean and standard deviation of the top 30 most frequent words was computed from the text corpus and the z-scores for all 12 texts were then computed. Then, taking each text in turn using it as an unknown against the other 11 texts, the delta scores were then computed. The results obtained using the Delta method were broadly similar. The results for Jefferson's early work, particularly pointed to a certain homogeneity of style.

2.2 Creating a Discrimination Word Vector

In the next phase of the study, we develop a word vector that is capable of accurately discriminating between Jefferson's and Paine's text. For this we used the 71 known texts by Jefferson written prior to the writing of the ADOI [8] and the 22 known texts written by Paine, also prior to the writing of the ADOI, [9]. We chose only texts written prior to the writing of the ADOI because Jefferson habitually quoted from the ADOI in his later writing. These texts were roughly equivalent in size forming approximately 54,000 words each. The texts were divided into blocks of 1000 words each. 32 blocks for each author were then used as a training set and the remainder used as a test set.

We rejected the use of Burrow's Delta method for two reasons: firstly it is primarily aimed at selecting the correct author from many and our study is aimed at selecting one of two authors, secondly, there are some methodological issues that are of concern. The Delta method relies on the Euclidean distance measure and it is well known that this is less effective over larger dimensional data sets. See for example [10,11] text mining applications that use large word vectors for document classification. We found a significant improvement in reliability with the Delta method after adapting it to use cosine similarity [12]. Although Hoover [13,14] reports improvements in the Delta method using larger word vectors, we were unable to replicate these improvements using the standard Euclidean distance measure, in fact our results suggest that performance of the Delta Method degrades with word vectors of size greater than about 200 words.

We felt that by simply choosing words by frequency particularly for very large word vectors was a coarse grained measure of style and better results could be achieved by selecting a word vector based on its ability to discriminate between two authors. The problem then becomes one of combinatorial optimisation as we need to select the optimal subset of words from the concordance listing.

We initially tried to evolve a discriminant function using a genetic algorithm, but then changed tactics following the completion of [15] in which subsets of words were chosen from the concordances of literary works by 4 authors to assess the degree to which different authors' texts exhibit clustering tendency. In a comparison of two texts by two different authors we measured the intra-cluster distance for the two clusters and compared this with the inter-cluster distance between the two clusters. These distance ratios were then compared using

³ Age of Reason was also written during a difficult period in Paine's life – he was at that time condemned to death in France.

different word vector subsets of the concordance. This technique allowed us to investigate the clustering tendency of small word vectors which appeared to show promise of revealing why authorial style can be measured by word frequency. For example, it was noticed that some word pairs that clustered well also appeared together in for example, complex prepositions.

In addition to the intra and inter cluster distance measures we also computed an n-dimensional complex hull minimum enclosing space which acted as a means of testing unknown points in the n-space for authorship. We noted that for n blocks of data and an m-word vector, that $n > m$. This had the effect of dividing the space into three regions:

- The area within author 1's cluster.
- The area within author 2's cluster.
- The area outside both clusters.

This idea was developed in our current study. The choice of a subset of words to form a word vector from the words that appear in the joint concordance of Paine's and Jefferson's writing is a combinatorial optimisation problem. In order to find a good solution, we elected to use a genetic algorithm. The representation used was a simple binary string for which each bit position represented the presence or absence of a word in the chosen word vector. Bit 1 represented the most frequent word in the concordance, bit 2 the second most frequent etc.

For example in a concordance consisting of {the, and, to, of, from} – bit pattern 10101 represents word vector {the, to, from}. An initial population size of 5000 was chosen. Because of the constraint that the word vector should contain fewer words than data points, the initial population was seeded with vectors of different sizes ranging from 2 to 30 words. We elected to apply a large fitness penalty to any vectors containing more than 30 words. This resulted in a sparse bit pattern as the concordance contained more than 6000 words. Initially, the words that constituted the word vector in the initial population were chosen randomly. However, it was discovered that this approach created a population with a very poor overall fitness. It was long been established in authorship attribution studies that high frequency words are better indicators of fitness than lower frequency words. A typical concordance contains up to 60% of hapaxes, which are of dubious value as indicators of authorship. In a comprehensive study [25] also suggests that the most successful authorship attribution methods are based on high frequency words.

We then changed the method of initial word selection by biasing initial word choice according to the frequency with which the word appeared – a technique that biased selection to high frequency words, but did not rule out the inclusion of lower frequency words. This produced far better results.

Members of the population were then evaluated according to their fitness. Fitness being measured by the ability of the word vector to separate out the training set into two distinct cluster regions and to maximise the inter-cluster distance. We also discouraged the growth of larger word vectors by penalising word vectors over size 30 because of the constraint imposed due to the necessity to use word vectors of lower dimensionality than the number of available data points.

We initially used a standard method of crossover. However, because of the sparse nature of the word vector, we introduced a hybrid operator called "grow". This operator created an extra word in the word vector (i.e. it turned on a bit). Additionally, it worked significantly better by carrying out local hill-climbing on

the grow operation in which the bits in the immediate area of the chosen bit (5 bits on either side) were tested for improvements to fitness. The resultant bit string was then added to the next generation.

Using this method, we found that optimal word vectors were created from vector sizes of between 25-30 words. As expected, these vectors were dominated by high frequency function words, but also contained a few words from the middle range of the Zipf distribution. Candidate word vectors were then tested on the test set for robustness.

Using this method, after several runs, we were able to create word vectors with completely disjoint clusters and with an accuracy of about 96% on the test set.

We were then ready for the next stage, which was to test the word vector on the ADOI itself.

2.3 Testing the Word Vector on the ADOI

In order to carry out an authorship study of ADOI, it is important to note the context in which it was written and also any sources that may have been used. Two texts in particular are important – *The Summary View of the Rights of America* written as *Draft of Instructions to the Virginia Delegates in Continental Congress* by Thomas Jefferson in 1774 and also *The Virginia Constitution*, thought to have been written in 1776 [8]. Nine textual similarities between SV and ADOI (i.e. passages that appear to have been directly taken or edited from SV) were identified and a further 16 similarities where passages in ADOI were clearly inspired by SV were also discovered. It was also noted that the 23 grievances against the King forming the central part of ADOI were either copied directly from, or edited from *The Virginia Constitution*⁴. Indeed 16 of these grievances appear in exactly the same order in both documents. It was therefore decided to analyse sections of the ADOI separately as it is clear that the central portion is either copied or edited from a text attributed to Jefferson. The ADOI was therefore subdivided into two parts:

- The List of Grievances (ADOIA)
- The Opening and Closing Statements of the Declaration (ADOIB).

We justify this on the grounds that the grievance list is copied or edited from *The Virginia Constitution*. ADOIB therefore contains 1104 words. ADOIA contained only 441 words. ADOIB was then tested using the word vector generated and tested over the Jefferson/Paine texts. The first word vector chosen was:

{ and, to, that, not, this, or, by, with, on, at, so, than, who, may, some, one, first, only, every, what, were, there, now, such, yet, same, when, out, had, up }

This word vector was then treated as a point in n-space and the position of this point in n-space was tested for inclusion within the Paine/Jefferson clusters defined using a convex hull algorithm for the minimum containing area of the cluster. The point was found to lie outside both the Paine and Jefferson

⁴ There has also been some speculation as to whether the Virginia Constitution really preceded The ADOI [3]. Other documents have also been suggested as source for the ADOI, e.g. The Mecklenburg Declaration, which was commented on by John Adams, though it was later shown to post-date the ADOI [16].

cluster. The distance of the point from the computed centroids and the minimum distance from the Paine/Jefferson clusters was then computed. The minimum distance ratios for the Paine/Jefferson clusters was 1: 10.1, for the centroids it was 1: 8.8 – demonstrating that the point was far closer to the Paine cluster.

This exercise was repeated for 19 other word vectors all of which had high scores. In three cases out of 20, the vector for ADOIB was found to be inside the Paine cluster, in the other 17 cases, it was closer to the Paine cluster, using minimum distance, by a ratio of approximately 9.7 : 1. While this is not completely conclusive, we suggest that it casts doubt on Jefferson’s authorship.

As a comparison, we applied Burrow’s delta method to ADOIB against each of the files used in the first part of the study [2]. In Table 1, the lowest scores indicate authorship – in this case suggesting Paine as a likely author – but not conclusively. For comparison, we also applied the test to *The Summary View of The Rights of America* (table 2).

Table 1 Delta Scores for The Declaration of Independence

	j1	j2	j3	j4	j5	j6
ADOIB	1.529	1.564	1.69	1.721	1.709	1.637
	auto	aor	comm	crs1	crs2	rom
ADOIB	1.647	1.682	1.254	1.202	1.248	1.502

Table 2 Delta Scores for *Summary View of the Rights of America*

	j1	j2	j3	j4	j5	j6
Summary View	0.961	0.839	0.923	1.046	0.971	0.981
	auto	aor	comm	crs1	crs2	rom
Summary View	1.058	1.51	1.375	1.477	1.41	1.411

The results from table 2 indicate Jefferson as likely author of *Summary View*.

2.4 Further Tests based on Vocabulary and Grammar

In order to investigate the authorship of ADOI further, we carried out the following tests:

1. A Test of vocabulary usage based on words appearing in ADOIB.
2. A Test of Phrasal usage based on fragments taken from ADOIB.
3. A Test of Grammatical Usage based on the Function Words And/To

And/To appear in all word vectors that were capable of discriminating between Jefferson/Paine with high reliability. They were also high frequency words that differed markedly in their frequency in texts attributed to Paine or Jefferson. So these were investigated in much greater detail.

ADOIB consists of 441 different words, of which two are alternate spellings of the word “independent” in The Adams’ version. Of these, 340 are hapaxes.

The study now concentrated on the set of texts by Jefferson/Paine written by them prior to the writing of ADOI. A search was conducted through all of these texts using all words that appeared in the ADOIB. The results, shown in Table 3 are not particularly conclusive either way.

Table 3 – Vocabulary Usage By Paine and Jefferson

	Jefferson	Paine	Both	Neither
Vocabulary	71	65	113	162
Hapaxes	38	41	98	143

For the next stage of the study, 344 phrases comprised of word collocations from ADOIB were constructed. The aim of this exercise was to determine the extent to which either author used phrasal fragments contained within ADOIB as well as the use of, for example, complex prepositions or prepositional verbs – something that might not show up in a word-based study. These fragments were then categorised as grammatical or content fragments. A search was conducted for the existence of the fragments in the works of Jefferson⁵ and Paine written prior to the ADOI. Grammar-based approaches to authorship have been used elsewhere, for example [17,18] and particularly in forensic linguistics studies, for example [19,20]. [2] also differentiates certain function words by use of a part-of-speech tagger, albeit on a rather ad hoc basis

Initially, no distinction between the fragment types was made. The results are given (Table 4) in four categories as before:

1. Fragments unique to Jefferson.
2. Fragments unique to Paine.
3. Fragments used by both.
4. Fragments used by neither.

Table 4 – Fragment Usage Categories

Jefferson Only	Paine Only	Both	Neither
40	28	69	207

The fragments are now subdivided according to content or grammatical function: the results are given in Table 5.

Table 5 – Fragment Usage by Content/Function Word

	Grammatical	Content
Jefferson	16	24
Paine	16	12
Both	58	11
Neither	30	177

The results presented in Table 5 show that Jefferson scores proportionately higher on content but Paine scores proportionately higher on grammatical fragments. However, the

⁵ The section of the Virginia Constitution dealing with Grievances was omitted because of its very close correlation with the central section of the ADOI.

distribution of the fragments unique to each author is also interesting.

ADOIA (the section of the ADOI consisting of the grievances) is considered next along with the equivalent section of the *Virginia Constitution* (Table 6).

Table 6 – Fragments Occurring in ADOIA

	VC	ADOIA
Jefferson	10	3
Paine	2	8
Both	8	6

13 fragments are unique to Jefferson and 10 are unique to Paine within ADOIA, these are proportionately about what would be expected, because ADOIA makes up about 30% of the total. However, a comparison of ADOIA and the *Virginia Constitution* reveals an interesting pattern: only two out of ten Paine fragments also appear in *Virginia Constitution*, whereas 10 out of the 13 of the Jefferson fragments also appear in *The Virginia Constitution*. The pattern indicates consistency of authorship for Jefferson for ADOIA and the VC. However, it also points to the influence of Paine in ADOIA, but not in the *Virginia Constitution* – possible evidence that ADOIA was edited by Paine based on the *Virginia Constitution*?

The software used in the search marked areas of the ADOI where matches occurred and the pattern of matches for Jefferson and Paine differed considerably. Jefferson's matches were clustered, whereas, Paine's were more or less evenly distributed throughout ADOIB and were less frequent in ADOIA. It was conjectured that this might be due to the fact that sections of the ADOI were edited from *Summary View*. The same search was then repeated having removed *Summary View* from the Jefferson text corpus. The results are given in Table 7.

Table 7 – Fragments Occurring in ADOI

	Grammatical	Content
Jefferson	16	24
Paine	16	12
Jefferson (minus Summary View)	8	14
Paine	16	12

The drop in both grammatical and content fragments clearly shows the influence of *Summary View* on ADOI – it also indicates that Jefferson's presence in ADOIB is in no small part due to *Summary View* being used as a source for ADOI. This in itself does not disprove Jefferson as author – but suggests an explanation as to why it is difficult to obtain a definite result on the authorship tests described above.

2.5 Jefferson and Paine's Use of "And" and "To"

Both Jefferson and Paine showed remarkable consistency with which they used high frequency words and it was noted that in particular, their usage of the third and fourth ranking words *and/to* was consistently different. It was also noted that the concordance of ADOI ranked the word *to* above *and* – a pattern consistent with Jefferson. However, if the central portion of ADOI (ADOIA) is removed, then this pattern reverses – though

only just – making it more consistent with Paine. This called for a closer examination of the use of these words within the ADOI.

The most frequent words used by Paine and Jefferson are listed in Table 8. Both Jefferson and Paine consistently used *the* and *of* most frequently. However the frequency of their third and fourth words was consistently reversed. Jefferson used *to* more than *and* – the order is reversed for Paine. These words always appeared in the word vectors derived above. This ordering reversed for Jefferson later in life. In any authorship study it is important to use not only at methods that are capable of discriminating between authors, but also to attempt to understand why those differences exist. In this section we work towards a grammar-based analysis of Jefferson/Paine attempting to match this with grammatical constructs used within ADOI.

Table 8 – Jefferson/Paine Most Frequent Words

Word	Jefferson /1000 words		Paine /1000 words	
	Mean	SD	Mean	SD
the	59.94	8.74	69.00	13.56
of	44.04	8.12	46.38	5.70
to	40.08	7.78	30.76	6.67
and	26.73	5.97	35.81	4.44

Attention then focused on the use of these words, because of their potential to discriminate between Jefferson and Paine. It was also decided to examine why this difference between the authors existed. Function words *and* and *to* have multiple uses [21]. A recent study on forensic data [22] identified 16 different categories of use for the function word *to* and 38 different uses for the function word *and*. *To* is primarily used in either a *to-infinitive clause* or as a *preposition*. *And* is used mostly in *clausal co-ordination* and as a *sentence connector*, although it also has other less frequently occurring uses.

The function word *and* is used in the following ways in ADOIB:

- As a connective with a following to-infinitive clause.
- To co-ordinate a binomial phrase.
- Use as a connective for phrases or sentences.
- Use as a complex connective.

To is used in the following ways in ADOIB:

- As a non-finite clause.
- In conjunction with an empty *it* clause with a non-finite clause.
- Beginning a sentence as a marker of an infinitive verb.
- As part of a complex connective *to which*.
- Various uses as a preposition, i.e. as part of a complex preposition, as a marker of a prepositional verb and as a simple preposition.

The uses of *and/to* were then used to construct a feature set to test whether the use of *and/to* matched Jefferson's or Paine's use of *and/to*.

Once again the texts written prior to the publication of ADOI were used. From the survey of usage of *and/to* in ADOIB, the following feature set was constructed:

- The use of *and* as a connective followed by a to-infinitive clause.
- *And* used in a co-ordinated binomial and phrase.

- The complex connective *and such*.
- *To* used as a non-finite clause.
- Constructions using the empty *it* subject with a non-finite clause.
- *A to infinitive* beginning a sentence.
- The use of the copula followed by a noun phrase.
- The complex connective *to which*.

And used as a simple connective and *to* used as a preposition were excluded from the feature set. Of the chosen features, it was discovered that two of them never occurred in either Jefferson or Paine's chosen writing. The remaining features were tabulated by help of custom written programs and a standard part-of-speech tagging program with manual checking to remove spurious matches. The results of this survey are given in Table 9. The row labelled J Total gives the total number of the feature found in Jefferson's writings. P Total is the total for Paine's writing. The frequency of co-ordinated binomial and phrases was so extraordinary that is worth a separate mention. Co-ordinated binomial phrases [23] pair words from all four major grammatical categories using *and/or* (for this study, only *and* is considered). *And* may co-ordinate noun and noun, e.g. *fish and chips*, verb and verb, e.g. *go and see*, adjective and adjective, e.g. *black and white* or even adverb and adverb, e.g. *slowly and deliberately*. It was noted that Thomas Paine used co-ordinated binomial and phrases with an unusually high frequency (10.9 instances per 1000 words). In modern English they occur with a frequency of 0.8 per 1000 words and a survey of 20 authors contemporary to Thomas Paine revealed that no other author used them with a frequency as high as Paine.

Five of the seven chosen features were distributed according to a poisson distribution. For this Mosteller and Wallace's classic study [24] was referred to. P-values and likelihood ratios were then computed (Table 9). The results of these also suggest Thomas Paine as the likely author.

Table 9 – Feature Set Chosen from the American Declaration of Independence

Feature	to-inf	empty-it + to inf	binomial and + to inf	and + to inf	binomial and
J Total	774	0	6	22	122
P Total	708	65	16	12	585
J Mean	15.6	0	0.06	0.41	2.3
P Mean	13.1	1.2	0.3	0.22	10.9
DOI Target	13	3	1	5	18

Feature	To-inf begins	to be +
J Total	9	27
P Total	15	44
J Mean	0.17	0.51
P Mean	0.28	0.82
DOI Target	1	1

Table 10 – P-values and Likelihood Ratios for the Poisson Distributions

Feature	Empty-it+ to inf	binomial and + to inf	and + to
P Value	$P \geq 3$	$P \geq 1$	$P \geq 5$
P val -J	0	0.0952	0.0001
P val -P	0.1203	0.2591	0
likelihood ratio J: P	1 : ∞	1 : 2.72	too small

Feature	To inf begins	to be +
P Value	$P \geq 1$	$P \geq 1$
P val -J	0.1813	0.3935
P val -P	0.2591	0.5507
likelihood ratio J: P	1 : 1.36	1 : 1.48

As the likelihood ratios show, once again the feature set based on the use of *and/to* suggests Paine as the likely author of ADOI.

Table 11 – Statistics For To-Infinitive Clauses and Binomial And Phrases

Feature	Statistic	Jefferson	Paine
to-infinitive	Total	774	708
	Mean	15.6	13.1
	DOI Target	13	13
	σ (standard deviation)	2.84	3.73
	No. of SDs from Target	<1	<1
	Range of Values	9-21	7-20
binomial and	Ranking Percentile	60-70	40-50
	Total	122	585
	Mean	2.3	10.9
	DOI Target	18	18
	σ (standard deviation)	1.58	4.64
	No. of SDs from Target	10	2
Range of Values	0-6	5-24	
Ranking Percentile	>100	80-90	

Table 11 presents the data for the two remaining features. The *to-infinitive* does not present strong evidence either way, but the *co-ordinated binomial and phrase* provides further evidence to support Paine, the mean and standard deviation for Jefferson is so low that it makes it highly implausible for him (more than 10 standard deviations away from the target) to be the author, on the other hand the data provides some additional evidence for Paine as the target figure of 18 for ADOI is plausible as it is within the 80-90th percentile and is within two standard deviations of his mean score.

3 SUMMARY AND CONCLUSIONS

It has been demonstrated that Thomas Jefferson and Thomas Paine had consistent but different styles of writing using both Delta scores and discriminant analysis. It is argued that a method of authorship attribution for two authors should be based on an accurate word vector that is capable of discriminating between the two candidate authors with a proven level of accuracy. The choice of words for use is subset of the concordance – making it

a combinatorial optimisation problem. A genetic algorithm with local hill-climbing was used to find a suitable word vector and found that word vectors between 20-30 words were perfectly adequate. This word vector was applied to a training set consisting of texts written by Jefferson/Paine prior to the writing of the ADOI. A test was also used to test reliability of the word vector. The test for authorship was based on using an n-dimensional convex hull algorithm, which created a minimum defining space for each author. Twenty different word vectors were then applied to the ADOI. In three cases, the ADOI was within the minimum defining region for Paine the remainder were far closer to the Paine cluster. Tests using phrase fragments, both content and grammatical also pointed at Paine as the potential author. Finally an examination of the grammatical use of *to* and *and* also suggest Paine as the more likely author.

The ADOI itself proved to be quite complex in that one substantial section was clearly copied and edited from another document and substantial sections of *Summary View* appear to have been edited into it. On the face of it, assuming the reliability of the tests used, there should be evidence of Jefferson's hand in its construction, but this is absent in our results.

However, it is clear that Jefferson's contribution to independence and the ADOI is considerable. It is suggested that Paine was asked or instructed to draft the framework of the ADOI by Jefferson, or by another member of the committee such as Franklin or Adams. It is almost certain that Paine was instructed to use material from both *Summary View* and *The Virginia Constitution*. While authorship attribution can never be entirely conclusive, the results presented provide a possible case for Paine's authorship.

Paine's commitment to independence is beyond doubt and he was in Philadelphia at the right time. His skills as a writer were well known to many members of the committee and he was known personally by Benjamin Franklin. He had already written anonymously and it would have been undesirable for an Englishman to have drafted the ADOI. There is also evidence that *The Rough Draft* was copied from an earlier version [1]. This adds plausibility that the ADOI was originally drafted by someone other than Thomas Jefferson, however, we have no doubt that The Adams' Copy was indeed drafted by John Adams.

It is also worth noting that some people have suggested that Thomas Paine must have been responsible for drafting the ADOI because of the inclusion of the infamous anti-slavery clause (omitted from the final version). This on its own is insufficient as references to anti-slavery are made in *Summary View*, though these appear to be more for economic than humanitarian reasons.

We would like to suggest that the findings of our research indicate that the possibility of Paine's hand in the drafting of the original version of ADOI and feel that this should be further explored.

ACKNOWLEDGEMENTS

We would like to thank Russell Gerrard for his comments on the statistical analysis and Sheila Muntun for her interest in the project and for persuading the British Library to lend us several old texts. We would also like to thank Jane Riedel and George Collins for their comments and would also like to thank Andrew Tuson for his moral support for this research.

REFERENCES

- [1] Hazelton, John H. The Declaration of Independence: It's History. Dodd, Mead & Co. (1906).
- [2] Burrows J. 'Delta': A measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing*, 17,3,2002 pp. 267-287. (2002).
- [3] Moody J. Thomas Paine: The Author of The Declaration of Independence. John Gray & Co. (1872).
- [4] Boyd Julian P. (1999) The Declaration of Independence: The Evolution of the Text. Ed. Gerard W. Galt. University Press of New England (1999).
- [5] Adams, Charles Francis. The Life and Works of John Adams. Reprinted AMS Press (1856).
- [6] Whissell C.
<http://www49thparallel.bham.ac.uk/back/issue9/whissell.htm>. Accessed June 2005
- [7] Thompson Bradley C. The Revolutionary Writings of John Adams. Liberty Fund, Indianapolis (2000).
- [8] Boyd Julian P. The Writings of Thomas Jefferson Vols. 1-27. Princeton University Press, Princeton, NJ. (1950).
- [9] Conway, Moncure D The Writings of Thomas Paine Vols. 1-4 Reprinted by Ayer and Co.. (1894).
- [10] Frigui H. and Nasraoui O. Simultaneous Clustering and Dynamic Keyword Weighting for Text Documents. In Berry M.W. (ed.) *Survey of text Mining: Clustering, Classification and Retrieval*. Springer (2003).
- [11] Korfhage R.R. *Information Storage and Retrieval*. Wiley, New York. (1977).
- [12] Aldridge W.E. The Burrows Delta Dilemma: Optimisation of Delta for Authorship Attribution. M.Sc. Thesis. City University, London. (2007).
- [13] Hoover D. Testing Burrows's Delta; *Literary and Linguistic Computing* 19,4, pp.453-475. Oxford University Press (2004).
- [14] Hoover D. Delta prime?; *Literary and Linguistic Computing* 19,4, pp. 477-495, Oxford University Press. (2004).
- [15] Smith P.W.H. The Clustering Tendency of Texts by Author. Unpublished. (2007).
- [16] Hoyt, William H. The Mecklenburg Declaration of Independence: A Study of Evidence showing that it is Spurious. Da Capo Press. 1972.
- [17] Baayen, R. H., van Halteren, H., & Tweedie, F. J. Outside The Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution. *Literary and Linguistic Computing*, 2, 110-120, (1996).
- [18] Lancashire, I. Phrasal repeats in Literary Stylistics: Shakespeare's Hamlet III.1. In S. Hockey & N. Ide (Eds.), *Research in Humanities Computing. Selected papers from the ALLC/ACH Conference Christ Church, Oxford*. Oxford: Clarendon Press. (1992).
- [19] Grant, T. D. Reviewing and Revising Stylometric Authorship Attribution for use in a Forensic Context. Paper presented at the International Association of Forensic Linguistics 5th Biennial Conference, University of Malta. (2001).
- [20] Smith Peter W.H. and De Jong G. Speaker Identification: Function Words and Beyond. Presented at The International Conference on Forensic Linguistics. Cardiff, July 2005. (2005).
- [21] Quirk R., Greenbaum S., Leech G., Svartvik J. *A Comprehensive Grammar of the English Language*. Longman, London. (1989).
- [22] Smith, Peter W.H. and De Jong G. Speaker Identification: Function Words and Beyond. Presented at The International Conference on Forensic Linguistics. Cardiff, July 2005. (2005).
- [23] Biber D., Johansson S., Leech G., Conrad S and Finnegan E. *Longman Grammar of Spoken and Written English*. Longman, London. (2002)
- [24] Mosteller F. and Wallace D.L. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Prentice Hall. (1964)
- [25] Grieve J. Quantitative Authorship Attribution: A History and an Evaluation of Techniques; Master of Arts Thesis, Department of Linguistics, Simon Fraser University (2005).