The *Federalist* Revisited: New Directions in Authorship Attribution

D. I. HOLMES and R. S. FORSYTH
Department of Mathematical Sciences,
University of the West of England, Bristol, UK

Abstract

The *Federalist Papers*, twelve of which are claimed by both Alexander Hamilton and James Madison, have long been used as a testing-ground for authorship attribution techniques despite the fact that the styles of Hamilton and Madison are unusually similar. This paper assesses the value of three novel stylometric techniques by applying them to the *Federalist* problem. The techniques examined are a multivariate approach to vocabulary richness, analysis of the frequencies of occurrence of sets of common high-frequency words, and use of a machine-learning package based on a 'genetic algorithm' to seek relational expressions characterizing authorial styles. All three approaches produce encouraging results to what is acknowledged to be a difficult problem.

1. The Case of the *Federalist Papers*

During 1787 and 1788, seventy-seven articles were printed in four of New York City's five newspapers, with the aim of persuading New Yorkers to support ratification of the proposed new constitution of the United States of America. These articles appeared under the pseudonym Publius and, as it happens, were unsuccessful: 56% of the citizens of New York state voted against ratifying the constitution. Undeterred by this setback, Publius re-issued these propaganda pieces in book form in May 1788, together with an additional eight essays that had not previously been published, so that delegates at the Constitutional Convention, then sitting, might be swayed by their case in favour of federalism. The New York delegation did eventually abandon opposition to the constitution, but mainly, it is thought, because nine of the thirteen states ratified, leaving New York potentially isolated (Wills, 1982). The book, however, has remained in print for over 200 years.

Speculation concerning the identity of Publius was widespread at the time, and gradually it became accepted that General Alexander Hamilton had been heavily involved in the composition of the *Federalist Papers* but that he had not written them all alone. Hamilton died in a duel with Aaron Burr in 1804, and in 1807 a Philadelphia periodical received a list, said to have been made by Hamilton just before his fatal duel, assigning specific papers to specific authors—himself, John Jay, and James Madison (the fourth president of the United States). Not until he retired from the presidency did Madison concern himself with asserting authorship of particular *Federalist* papers, but in 1818 he claimed to have written numbers 49–58 as well as 62 and 63—papers that Hamilton's list had ascribed to himself. Thus twelve of the eighty-five papers were claimed by both Hamilton and Madison.

From that time till 1964, when Mosteller and Wallace published the first edition of their book, *Inference and Disputed Authorship: the Federalist*, scholarly opinion was divided. Both Hamiltonian and Madisonian authorship of the twelve disputed papers were seriously argued on a variety of historical and stylistic grounds. For reference, the situation that Mosteller and Wallace found, with regard to authorship of particular numbers of *The Federalist* is tabulated in Table 1.

Table 1 Authorship of the *Federalist Papers*

Paper Number	Author
1	Hamilton
2–5	Jay
6–9	Hamilton
10	Madison
11–13	Hamilton
14	Madison
15–17	Hamilton
18–20	Joint: Hamilton and Madison
21–36	Hamilton
37–48	Madison
49–58	**Disputed**
59–61	Hamilton
62–63	**Disputed**
64	Jay
65–85	Hamilton

In the (extended) 1984 edition of their monograph, Mosteller and Wallace report six separate studies bearing on the question of who wrote these twelve disputed papers:

(i) a preliminary study involving a linear discriminant function;
(ii) 'the main study';
(iii) various sensitivity analyses of the main study;
(iv) a 'weight-rate analysis';
(v) a 'robust hand-calculated Bayesian analysis';
(vi) 'three-category analysis'.

They also describe some other subsidiary investigations of *Federalist* and other themes; but we concentrate here on four of the items above, particularly on what they call 'the main study'.

1.1 *A Pilot Study*

Unpublished work by Frederick Williams and Mosteller had convinced them that sentence length was a poor discriminator. Indeed Hamilton and Madison are

Correspondence: D. I. Holmes, Department of Mathematical Sciences, University of the West of England, Frenchay Campus, Coldharbour Lane, Bristol BS16 1QY, UK.

Table 2 Sentence length statistics

	Hamilton	Madison
Mean sentence-length (words)	34.55	34.59
Standard deviation (words)	19.2	20.3

almost identical on this measure, as can be seen in Table 2.

A glimmer of success came in a pilot study using a linear discriminant function based on the following variables: proportion of nouns; proportion of adjectives; number of one-letter and two-letter words; frequency of the definite article. A discriminant function derived from undisputed papers by Hamilton and Madison assigned nine of the disputed papers to Madison and three to Hamilton.

This result was suggestive rather than convincing, so Mosteller and Wallace turned their attention to what they call 'marker words', having been informed that 'while' and 'whilst' were used with very different frequencies by the two authors and having noticed that 'upon' and 'on' behaved in a similar fashion.

Table 3 Hamiltonian and Madisonian marker words

	Rates per Thousand Words			
	On	Upon	While	Whilst
Hamilton	3.28	3.35	0.28	0.00
Madison	7.83	0.14	0.02	0.48

1.2 *The Main Study*

In their main study, Mosteller and Wallace used Bayes' Theorem to draw inferences about the probabilities of the competing hypotheses that Hamilton or Madison wrote the disputed *Federalist Papers*, based on evidence provided by the rates of usage of thirty marker words, including four shown in Table 3. They concluded that Madison was very likely the author of all twleve disputed papers. This was a considerable intellectual feat, requiring much ingenuity as well as hard work. First, they collected a large amount of text from undisputed works by both authors, 94,000 words written by Hamilton and 114,000 words by Madison. It should be noted that some of these writings were from outside *The Federalist*: less than 10,000 words of their Hamilton sample were from non-Federalist sources but in Madison's case the majority, almost 73,000 words, was from his other works, composed over a 25-year period. This fact necessitated a subsidiary study of Madison's literary output to make sure that his rate of usage of the kind of words being counted (mostly high-frequency function words such as 'any', 'by', 'from', 'there' and so on) was indeed stable over time.

The next step involved choosing suitable marker words. This was achieved in several sweeps through subsets of the texts, which fell into two main phases. Firstly, well over 300 potential marker words were tested against screening texts of about 40,000 words by each author. Only 165 words, that showed significant ability to discriminate between texts by Hamilton and Madison, were retained. Then these 165 were whittled down to the final thirty by testing on a second set of texts that had not been used in the initial screening.

At this point Mosteller and Wallace had their discriminators and could apply Bayes' Rule to make statistical inferences. This rule for amending probabilities in the light of evidence (Bayes, 1763) lies at the heart of the main study. Mosteller and Wallace used logarithms of odds rather than probabilities and took advantage of the fact that they were working with only two possibilities, so they were able to work with a simplified version of Bayes' Rule that can be expressed as:

Final log-odds = initial log-odds + log-likelihood

Log-likelihoods for each marker word, derived from differential rates of usage, were simply added (assuming independence between words) to reach a final assessment.

However, it is perhaps not generally appreciated that Mosteller and Wallace carried out in this study *two logically separate stages* within the overall Bayesian reasoning process. The first stage—arriving at appropriate log-likelihoods given particular word counts in papers of known origin—involved much greater mathematical complexity than the second—applying those log-likelihoods to decide the authorship question.

The need for an initial stage arose because even with quite large amounts of text and relatively common words, the observed mean rates of word usage were insufficiently accurate to be used as they were. This can be illustrated by the word 'also', which occurred at a rate of 0.28 per 1000 words in Hamilton's sample, that is twenty-six times in 94,000 words. While 94,000 words is a large sample, twenty-six occurrences is not; and the standard error of the mean value 0.28 comes to 0.055 which is by no means negligible.

So Mosteller and Wallace studied the behaviour of ninety function words not chosen in the final pool of thirty markers to arrive at a realistic mathematical model of word usage behaviour. They concluded that, with few exceptions, word frequencies were badly modelled by assuming a Poisson distribution but that the negative binomial distribution gave good fits for the sort of words they were interested in. However each negative binomial model for each of thirty marker words required four parameters to be specified and in most cases there was not enough text simply to estimate these values from the data—as explained above. To surmount this problem they treated these parameters as uncertain variables and applied Bayes' Rule to compute a 'posterior' estimate of their values using texts of known authorship. Of course, this in turn required a model of the prior distribution for such values: their investigation of the ninety non-marker words convinced Mosteller and Wallace that the Beta distribution was satisfactory for this purpose.

The practical effect of this preliminary Bayesian stage before the main authorship analysis was to moderate the difference between Hamiltonian and Madisonian markers, especially for relatively rare words. This they interpreted as giving further protection, additional to the double screening of potential markers, against any bias introduced by selecting only highly discriminatory words.

At the end of all this they were able to combine the

evidence from thirty separate marker words and thus derive posterior log-odds in favour of Madison's authorship of the disputed papers. The weakest of the twelve resulting log-odds were those for paper 55, representing a multiplier of about 240 in favour of Madison, sufficient to overturn any but the most extreme prior odds on Hamilton.

1.3. *A Weight-Rate Analysis*

This analysis can be seen as an extension of the pilot study using a linear discriminant function (Section 1.1) in the light of the main study's findings. It was performed largely to find out whether a classical, i.e. non-Bayesian, statistical analysis would support the conclusions of the main study.

Firstly a pool of 117 potentially discriminating words was tested against a screening set of twenty-three papers by Hamilton and twenty-five by Madison. Only the most discriminatory twenty words were retained. Of these, the best ten, in order of importance are shown in Table 4. Here a plus sign indicates a word more used by Hamilton, a minus sign a word more used by Madison.

Table 4 The best ten discriminatory words

'upon'	+
'whilst'	−
'there'	+
'on'	−
'while'	+
'vigor'	+
'by'	−
'consequently'	−
'would'	+
'voice'	−

For each of the twenty selected marker words a weight was calculated to form a linear discriminant function. This function could be applied to a paper of known or unknown authorship by multiplying the rate of occurrence of the twenty markers (scaled per 1000 words) each by its weight and summing these products to give a numeric value y', which was positive for Hamilton and negative for Madison. Actual numeric values of y' were arbitrary; unlike log-odds, they were merely indicators.

On the screening set, the mean value of y' was 0.87 for twenty-three Hamilton papers and −0.41 for twenty-five Madison papers. As these papers were used to select words and their weights, this constitutes a self-test. On a calibrating set of unseen papers, the mean values of y' were 0.92 for twenty-five Hamilton papers and −0.38 on twenty-five Madison papers. It turned out that there was no overlap on the screening set or, more importantly, on the calibrating set: the lowest-scoring Hamilton paper and the highest-scoring Madison paper in the calibrating set were about half a standard deviation apart; and although the standard deviation of the y' values increased from screening to calibrating sets, the mean y' values of the two authors were still approximately 4.5 standard deviations apart.

Based on the means and standard deviations of y' values calculated from this calibration set, all twelve disputed papers except number 55 were outside the 99% confidence interval for Hamilton, while all but numbers 55 and 56 were inside the 90% confidence interval for Madison. Thus a non-Bayesian statistical analysis gave very similar results to the main, Bayesian, study.

1.4 *The Robust Bayesian Analysis*

In this study Mosteller and Wallace sacrificed information in order to avoid some of the difficulties connected with choosing a suitable family of prior distributions and estimating what have come to be known as 'hyper-parameters' for such distributions.

The problem of modelling word usages was simplified by dichotomizing rates of usage: common words were categorized into two classes, above or below the median rate (for both authors combined); less common words were also put into two classes, zero or non-zero rate of occurrence in a 2000-word block. Thus the information about each word could be encapsulated in a 2 × 2 table. (The advantage with this simplification was that the whole study could be executed using only slide rules.) An example for the word 'to' appears in Table 5.

Table 5 Four-fold frequency table for word 'to'

Rates for 'to' (compared to cutoff rate of 37.805 per thousand words)		
	Low	High
Hamilton	7	16
Madison	16	7

These figures were obtained on a screening set of forty-six papers, twenty-three by each author, chosen to be close to 2000 words in length. In fact the range of paper lengths was from 1728 to 2282 words. Table 5 shows that in seven of his twenty-three papers Hamilton used the word 'to' less than 37.805 times per 1000 words and in sixteen papers more often than that — *vice versa* for Madison.

Mosteller and Wallace tried 193 words on this screening set and whittled this pool down to thirty-one that showed ability to discriminate between Hamilton and Madison. For each of these thirty-one words, a 2 × 2 table similar to that above was obtained and used to compute appropriate likelihood factors.

To test the procedure, Mosteller and Wallace used the thirty-one tables to calculate posterior log-odds on Hamilton's authorship (assuming even prior odds between Hamilton and Madison) for each of the forty-six screening papers. The mean posterior log-odds for the twenty-three Hamilton papers came to 13.95 and for the twenty-three Madison papers it came to −14.25. On a validation set of thirty-one papers, not used to form the 2 × 2 tables, the mean log-odds were 10.2 for thirteen Hamilton papers and −8.2 for eighteen Madison papers. Thus, as expected, there was a shrinkage from screening to calibrating papers; but it is worth noting that even log-odds of 8.2 represent odds of 3641 to 1 in favour of the correct author.

When applied to the twelve disputed papers, this method gave mean posterior log-odds of −5.66. All were strongly Madisonian except number 52 (posterior

odds of 7 to 1 in favour of Madison) and number 55 (once again) where the resulting odds were 10 to 1 in favour of Hamilton.

1.5 *Appraisal*

In an attempt to evaluate the contribution of Mosteller and Wallace, first it is worth noting their own self-assessment (Mosteller and Wallace, 1984):

> We tracked the problems of Bayesian analysis to their lair and solved the problem of the disputed Federalist papers.

As this ordering suggests, they were primarily interested in demonstrating a practical application of Bayesian statistics, and in solving some of the problems associated with doing so, and only secondarily in settling the question of who wrote the disputed Federalist papers.

A second point worth making at this stage is that they have been more admired than emulated. Subsequent researchers have been reluctant to follow down the path they pioneered; and Bayesian reasoning has not played a major part in stylometric studies in the three decades since 1964. Indeed it is not much of an exaggeration to say that stylometrists have dropped the baton handed on by Mosteller and Wallace. The fact that Mosteller and Wallace obtained rather similar results with a non-Bayesian approach may have told against their advocacy of Bayesianism.

From a polemical point of view, Mosteller and Wallace may thus be said to have failed (rather like Hamilton and Madison before them!) in their main objective: they have impressed but not persuaded the majority of their intended audience. They might actually have had more followers if their work had been less comprehensive, for example if they had never published their book but only a paper on the robust Bayesian analysis, which is easier to understand and copy than the main study.

Nevertheless, although Bayesian reasoning has virtually disappeared from stylometric studies, the *Federalist Papers* re-appear from time-to-time usually as a testing ground for attribution techniques. They represent a severe choice as test problem since Mosteller and Wallace (1984) do caution their readers:

> Hamilton's and Madison's styles are unusally similar; new problems, with two authors as candidates, should be easier than distinguishing between Hamilton and Madison.

An example study using the *Federalist Papers* as a test case is that of McColly and Weier (1983) who advocate a likelihood-ratio approach to attribution based on the assumption that the frequency of occurrence of a particular marker word is distributed as a Poisson random variable. Their study was directed at the Middle English *Pearl*-poems, but to validate the underlying assumptions of their methodology they turned to the *Federalist Papers*. Six essays from the *Federalist* corpus were selected, two of Hamilton (numbers 60 and 66), two of Madison (numbers 14 and 45) and two of the disputed group (numbers 53 and 57). Thirteen of the thirty Mosteller and Wallace marker words were chosen, words which occurred once per 1000 in at least one essay, and the likelihood ratio test on these thirteen variables was applied to all pairs of essays. The resulting Chi-square values seemed to confirm Mosteller and Wallace's findings that both disputed papers were written by Madison. A larger analysis was then conducted on a pool of sixty-four content-independent function words with much less satisfying results, the null hypothesis of, in literary terms, common authorship, being rejected in all paired cases except for the two disputed papers. In statistical terms the likelihood ratio test depends upon a null hypothesis of equal Poisson parameters for the pool of words under study, and Damerau (1975) has provided evidence to suggest that this assumption is unrealistic.

The behaviour of one of Morton's 'proportionate pairs' (Morton, 1978), in particular the pair *on* / (*on* + *upon*) was investigated in the context of the *Federalist Papers* by Merriam (1989). Merriam grouped the individual *Federalist Papers* in sets of three in order to get convenient sample sizes of about 6000 words and found that Hamilton and Madison have internally consistent usage patterns, which differ from each other by amounts in excess of chance variation.

A recent study on the *Federalist* problem has been conducted by Kjell (1994) who used letter-pair frequencies as input to a neural network, which was trained to discriminate between Hamilton and Madison using back-propogation. His results broadly confirm those of Tweedie *et al.* (1994) and Mosteller and Wallace.

In this paper we will assess the value of three novel stylometric techniques for authorship attribution by applying them, somewhat bravely in view of Mosteller and Wallace's warning, to the *Federalist* problem.

2. Vocabulary Richness

A multivariate approach to measuring the vocabulary richness of a literary text has been proposed by Holmes (1991, 1992) and successfully applied to the corpus of Mormon scripture. The approach highlighted a similarity/dissimilarity structure amongst the textual samples studied. The five variables involved each measured the richness of vocabulary of a writer in some sense, yet were stable with regard to text length—a most important attribute. Denoting the text length of N words, the number of different words in the text by V, and the number of words used exactly r times in the text by V_r ($r = 1,2,\ldots$), then the variables employed were:

(i) $$R = \frac{(100 \log N)}{\left(1 - \left(\frac{V_1}{V}\right)\right)}$$

(ii) $$\frac{V_2}{V}$$

(iii) $$K = \frac{10^4(\Sigma_{i=1}^{\infty} i^2 V_i - N)}{N^2}$$

(iv) and (v) Parameters α and θ, the standardized values of the parameters from the Sichel (1975) distribution, a theoretical model for vocabulary distributions, V_i. The slope of the head of the distribution is defined by α, whilst that of the tail by θ.

As recommended by Sichel, parameters α and θ were computed from the observed vocabulary distributions for nouns only, and it was more practicable also, to evaluate K from the noun distributions. These five variables, therefore, collectively cover the whole structure of a writer's vocabulary frequency distribution.

To these we now propose to add a sixth variable, the W index proposed by Brunet (1978) as a measure of vocabulary richness. The index is defined by

$$W = N^{V^{-a}}$$

where a is a constant in the range 0.165 to 0.172. Here we use $a = 0.170$. Brunet claims that values of W are relatively unaffected by text length and that it is author-specific.

To apply vocabulary richness techniques to the *Federalist* problem a sample of thirty-three papers was shown. All twelve Disputed papers were selected, along with the three Joint papers, all five Jay papers, a random sample of eight Hamilton papers (numbers 7, 11, 21, 24, 31, 36, 70, and 72) and a random sample of five Madison papers (numbers 10, 39, 40, 42, and 44). All thirty-three texts were run on the Oxford Concordance Program, after which from each word list nouns were manually tagged, lemmatized (singular and plural forms subsumed) and extracted into separate vocabulary distributions. The thirty-three distributions obtained were arranged in identical groupings for comparability and the SICHEL computer program (see Holmes, 1992) was then run on each resultant data set. The values of the six vocabulary richness variables were readily obtained from both the OCP and SICHEL outputs.

As a check against any dependence on text length, the six variables were correlated against text length N. The only significant correlation coefficient obtained was that between V_2/V and N (0.449). Although Sichel (1986) shows stability of the proportion of hapax dislegomena within an author for $1000 < N < 400{,}000$, we may be too near the lower end of this range with the *Federalist* papers as compared with the Mormon texts previously examined where N was around 10,000. It was accordingly decided to omit variable V_2/V from the analysis. One interesting result was that the correlation coefficient between the type-token ratio (V/N) and N was -0.705, confirming the fact that any stylometric study using this ratio should always work with equal size text lengths. Table 6 lists the values of N, α, θ, K, R and W for the textual samples.

The first objective of this analysis must be to find the natural groupings, if any, amongst the textual samples. Statistical grouping techniques such as cluster analysis seek to form 'clusters' or 'groups' of individuals such that individuals within a cluster are more similar in some sense than individuals from other clusters. The result may be presented in the form of a dendrogram, a two-dimensional diagram illustrating the mergers which have been made at each stage of the procedure. Figure 1 shows the dendrogram computed by applying an average-linkage cluster analysis to the twenty-one *Federalist* samples obtained by excluding, temporarily, the disputed papers. The five-variable vocabulary

Table 6 Values of stylometric variables for the thirty-three sample papers

Paper	N	α	θ	K	R	W
Jay 2	1670	1.71336	0.991174	91.046	1015.20	11.9480
Jay 3	1453	2.23551	0.995272	140.165	932.40	12.4043
Jay 4	1642	4.11764	0.964783	103.049	1025.00	12.3315
Jay 5	1337	4.37702	0.975129	81.207	907.80	12.0763
Jay 64	2302	3.57756	0.991824	87.621	954.20	12.6961
Joint 18	2084	2.10943	0.981688	51.553	1094.11	11.7586
Joint 19	2020	0.00413	0.989795	45.130	1174.60	11.7295
Joint 20	1512	1.79164	0.979805	49.778	1177.30	11.4232
Disputed 49	1644	0.15388	0.995129	96.048	953.00	12.4386
Disputed 50	1103	2.44909	0.991781	81.240	948.80	12.2805
Disputed 51	1913	4.61325	0.989298	120.523	828.20	12.9292
Disputed 52	1841	1.25559	0.995235	110.719	831.64	12.9074
Disputed 53	2161	2.94301	0.993891	102.578	811.86	12.9961
Disputed 54	1997	4.80203	0.989639	180.071	834.98	13.4701
Disputed 55	2033	2.91582	0.989251	91.641	904.79	12.6734
Disputed 56	1560	0.43042	0.996872	184.553	797.40	13.3411
Disputed 57	2201	2.84495	0.989621	96.127	964.38	12.6915
Disputed 58	1341	1.30607	0.990314	114.150	1004.76	12.5851
Disputed 62	2380	2.81267	0.983281	74.302	931.30	12.3686
Disputed 63	3030	3.80748	0.983614	100.058	886.60	12.5996
Hamilton 7	2300	1.40744	0.985550	74.741	1128.50	11.9396
Hamilton 11	2512	0.93997	0.986407	49.934	1049.00	12.0991
Hamilton 24	1818	0.17177	0.990587	42.560	1141.20	11.9826
Hamilton 72	2035	3.89236	0.970456	58.230	1055.40	12.2677
Hamilton 36	2725	2.18118	0.986708	85.220	1005.72	12.6100
Hamilton 70	3062	0.48370	0.990039	42.500	977.57	12.3955
Hamilton 21	1994	0.86927	0.989103	65.441	1116.22	12.0700
Hamilton 31	1725	1.22874	0.994907	69.849	1079.52	12.2700
Madison 39	2598	3.87956	0.993313	157.430	822.14	13.8895
Madison 40	3019	0.85702	0.994907	131.281	951.64	13.0621
Madison 42	2772	2.77806	0.989262	124.702	902.45	12.8134
Madison 44	2897	3.88882	0.983759	152.853	878.29	12.9965
Madison 10	2996	3.50382	0.986728	81.810	887.52	12.7514

richness measures produce three main clusters; (i) the Madison papers plus Jay 3, Jay 64 and Hamilton 36, (ii) papers Jay 4, Jay 5 and Hamilton 72, and (iii) the remaining Hamilton papers, all three Joint papers and Jay 2.

Figure 2 shows the dendrogram for the Hamilton, Madison and Disputed papers subset. Here we notice that six of the eight Hamilton papers rapidly group, the Disputed papers intermingle with the Madisons and Hamilton 36, whilst Hamilton 72 appears as an outlier.

The dendrograms do not provide information concerning how much each of the five vocabulary richness

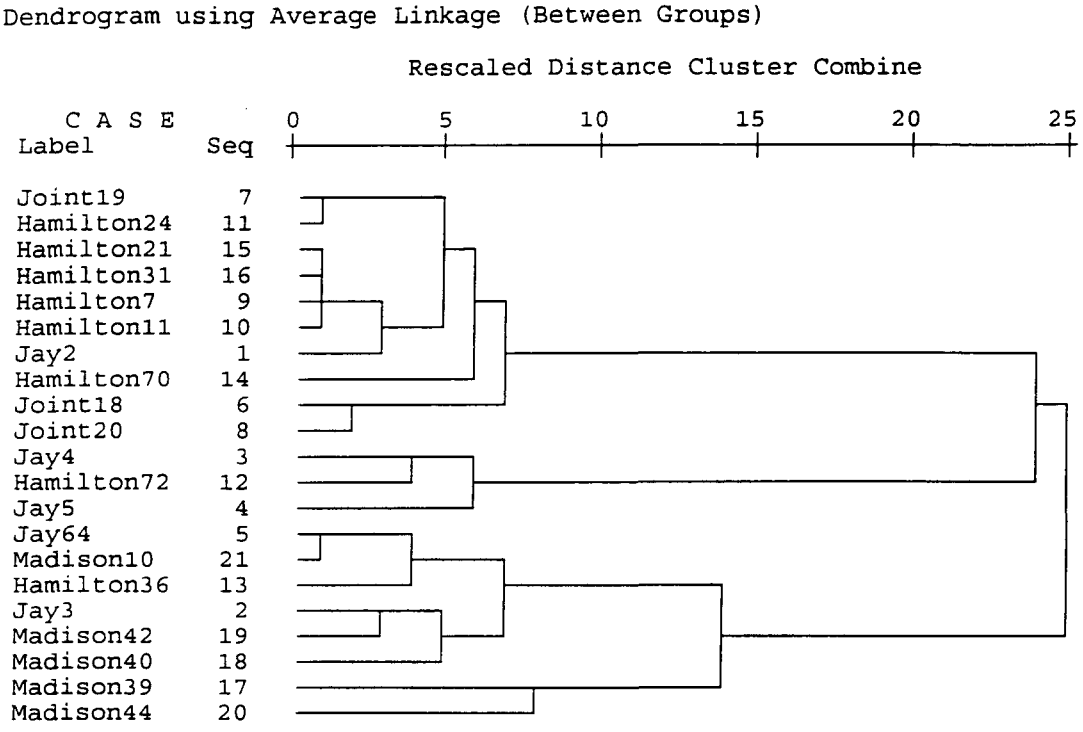

Fig. 1 Dendrogram excluding Disputed papers.

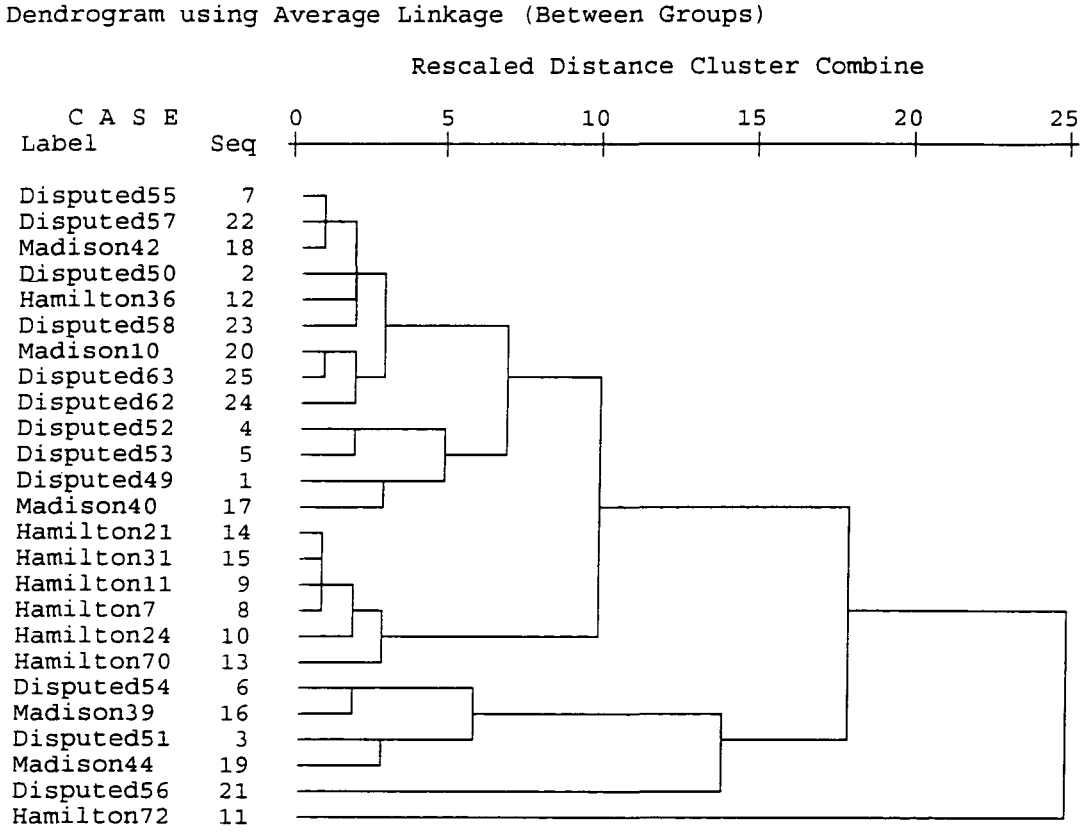

Fig. 2 Dendrogram including Disputed papers.

variables is contributing to these group differences. To answer this question we may turn to 'principal components analysis', which aims to transform the observed variables to a new set of variables which are uncorrelated and arranged in decreasing order of importance. The principal aim here is to reduce the dimensionality of the problem and to find new variables which will help to make the data easier to understand. These new variables (or components) are linear combinations of the original variables and it is hoped that the first few components will account for most of the variation in the original data. A principal components analysis was initially carried out on the (13 × 5) data matrix comprising only the Hamilton and Madison papers, the variables being scaled to possess zero mean and unit variance.

Figure 3 shows the thirteen papers plotted in the space of the first two principal components, which together account for 93.2% of the total variation and therefore provide a plot which conveys well the structure present in the original five dimensions. The Madison papers group well, as do the Hamiltons with the now familiar outlier Hamilton 72. It is the first principal component which is separating the Hamilton and Madison papers. This component has eigenvector

$$\begin{bmatrix} \alpha & 0.409 \\ \theta & 0.113 \\ K & 0.509 \\ R & -0.524 \\ W & 0.535 \end{bmatrix}$$

and therefore contrasts R against α, K and W. With high values of R and low values of K and W indicating a 'rich' vocabulary, we can see that the Hamilton papers are at the 'richer' end of the vocabulary spectrum.

Figure 4 now incorporates the Jay (J) and Joint (C) papers into the space of the first two principal components, which together account for 90.0% of the total variation. The three Joint papers group well and interestingly lie at the richest end of the first principal component axis. This raises the question of whether collaborative texts are always richer in vocabulary than texts from separate contributors.

Figure 5 looks at the equivalent two-dimensional plot for the Hamilton, Madison, and Disputed papers only, which accounts for 89.9% of the total variation in the five-dimensional data set. The grouping structure highlighted by Figure 2 is again evident, with the Disputed papers lying on the Madison side of the plot.

In all three principal components plots, the second principal component, which reveals Hamilton 72 as an outlier, contrasts α and θ. Hamilton 72 has an unusually high α value, hence its position in the multivariate plot.

The final analysis on the five-variable vocabulary richness measures employed a linear discriminant analysis on the Hamilton and Madison papers, followed by a prediction of group membership for each of the twelve Disputed papers. Using cross-validation, eleven out of the thirteen Hamilton and Madison papers were correctly assigned to their true groups. Cross-validation omits the first observation from the data set, develops a classification function using the remaining observations, then classifies the first observation. Next, it returns the first observation to the data set, omits the second observation, and repeats the same process. In the predictive part of the discriminant analysis, all twelve Disputed papers were assigned to the Madison group with the lowest probability of group membership being 87.9% for Disputed 58. This is an impressive result bearing in mind the Mosteller and Wallace findings. The vocabulary richness variables would

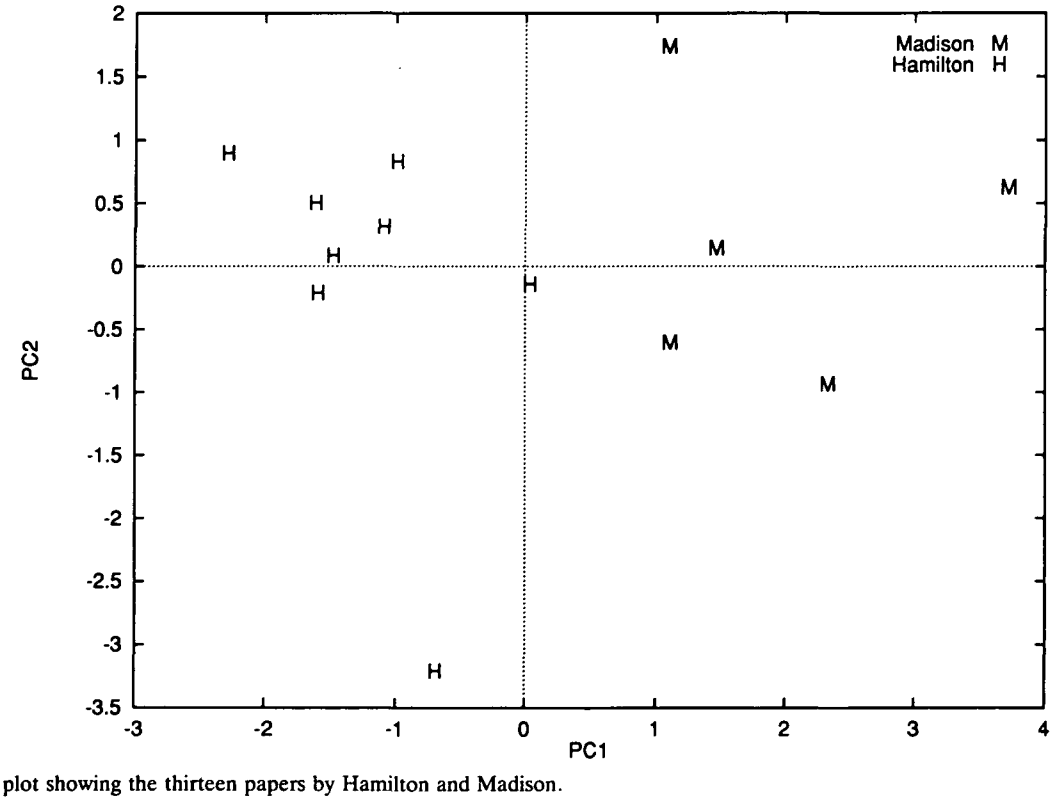

Fig. 3 PCA plot showing the thirteen papers by Hamilton and Madison.

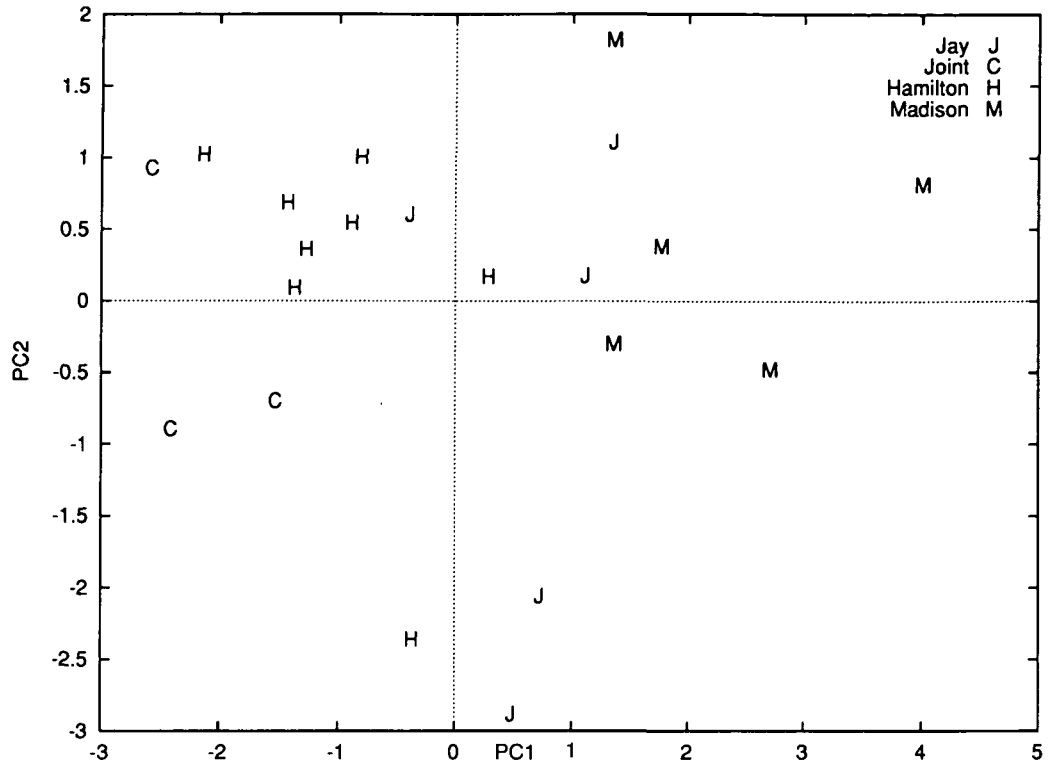

Fig. 4 PCA plot including the Jay and Joint papers.

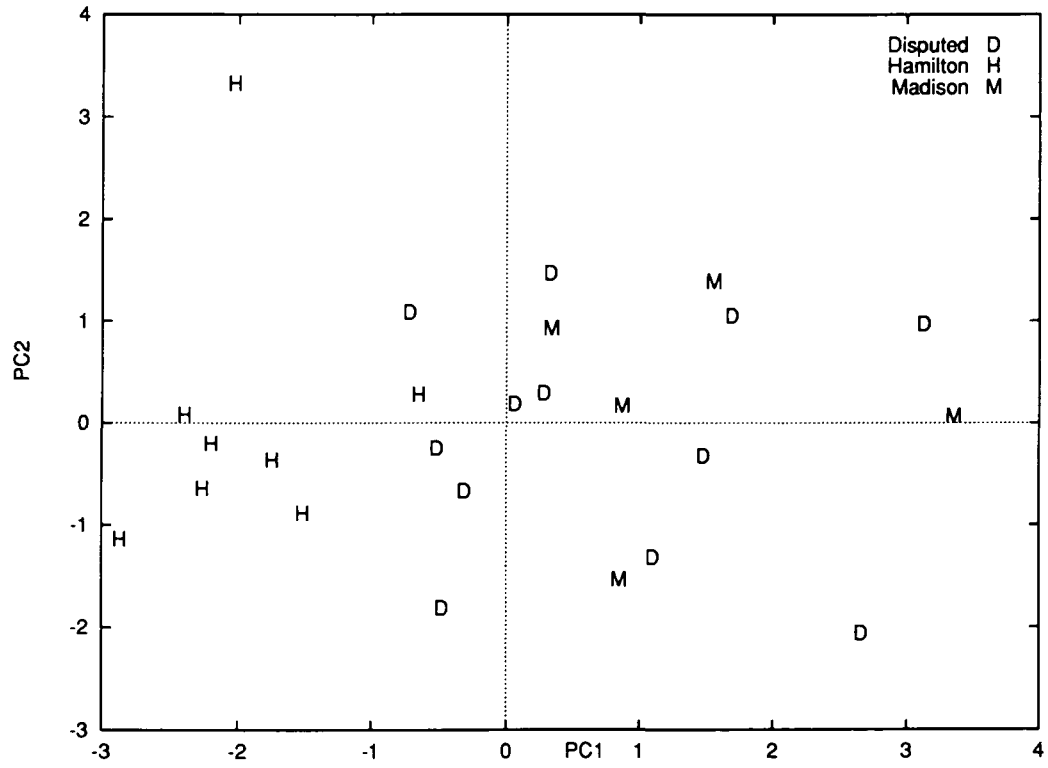

Fig. 5 PCA plot including the Disputed papers.

seem to be a good set of discriminators in a collective sense.

3. Word Frequency Analysis

In a ground-breaking study which we believe sets new directions for stylometry, Burrows (1992) employs what is essentially principal components analysis on the frequencies of occurrence of sets (typically 50) of common high-frequency words. His study was applied to a wide variety of authors and genres and achieved remarkable results as regards the clustering patterns of

textual samples in the space of the first two principal components. We used this technique on the *Federalist Papers*, using the same samples except for the random omission of three Hamilton papers (numbers 21, 31, and 36) to reduce the burden of computation and to balance the genuine Hamilton and Madison papers at five each.

Two separate investigations were conducted. The first used a set of forty-nine words from among the list of ninety high and medium-frequency words used as an initial pool by Mosteller and Wallace. Many of these ninety words appear only rarely in some of the papers, hence the reduction to forty-nine most commonly occurring high-frequency function words. This procedure produces a list almost the same as would be obtained by Burrows' method of choosing the fifty most common words. Table 7 shows these forty-nine high-frequency words. A large spreadsheet was compiled of the (49 × 30) word occurrences and these were standardized for all the thirty papers by computing occurrence rates per 1000 words.

The second investigation proceeded in a similar manner but exclusively used only the thirty Mosteller and Wallace marker words listed in Table 8.

Table 7 Forty-nine high-frequency words

a	by	is	one	this
all	can	it	only	to
an	every	its	or	was
and	for	may	so	were
are	from	more	some	which
as	has	must	such	who
at	have	no	than	will
be	if	not	that	with
been	in	of	the	would
but	into	on	their	

Table 8 The thirty Mosteller and Wallace marker words

according	considerable(ly)	probability
also	direction	there
although	enough	this
always	innovation	though
an	kind	to
apt	language	upon
both	matter(s)	vigorous
by	of	while
commonly	on	whilst
consequently	particularly	work(s)

Figure 6 shows the Hamilton and Madison papers in the space of the first two principal components for the forty-nine word data set. This plot accounts for 74.6% of the variation in the original forty-nine-dimensional matrix and the first principal component contrasts the use of 'the' against 'to', 'would' and 'a'. The Madison papers cluster well but are joined by Hamilton 70. Use of the thirty marker word data set in the space of the first two principal components produces the scatter plot shown in Figure 7.

If we now incorporate the Jay papers and the Joint papers, we obtain the principal components plot shown in Figure 8 for the forty-nine-word set, which accounts for 69.2% of the total variation. This is a remarkable plot, with clear clusterings evident for the Madison, Joint and Jay papers. The first principal component contrasts usage of 'the' against 'and' and 'would', whilst the second principal component contrasts 'and' against 'of' and 'a'. This impressive result adds weight to

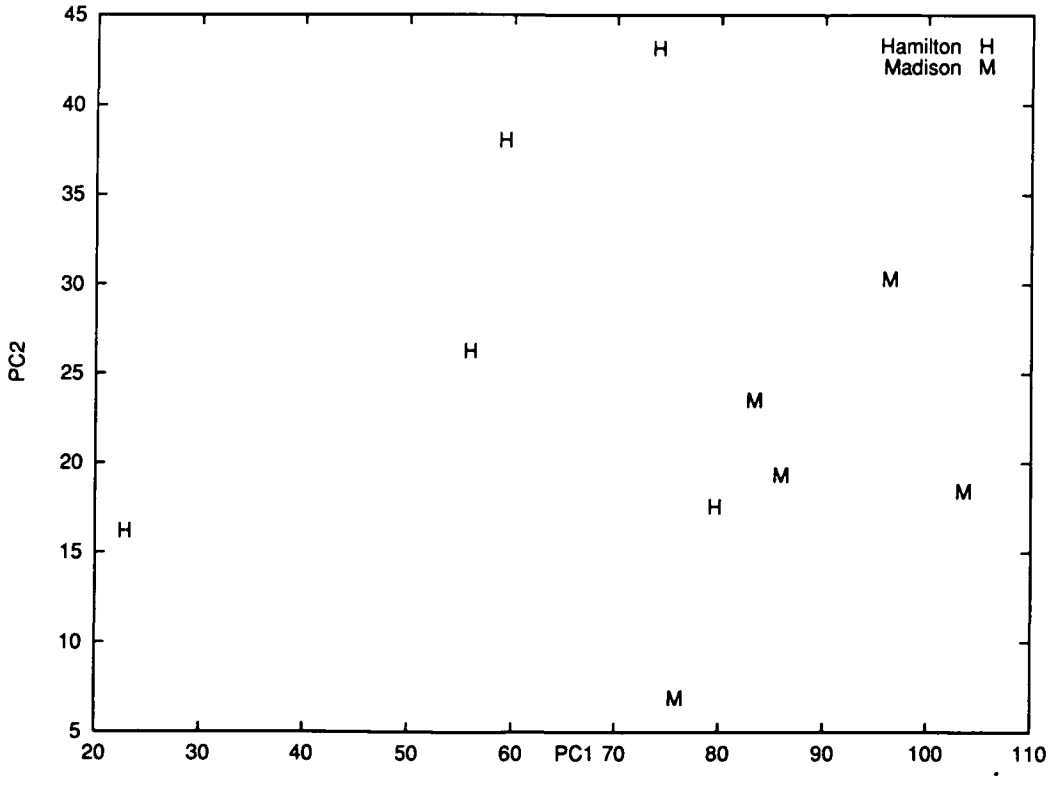

Fig. 6 Hamilton and Madison papers with forty-nine words.

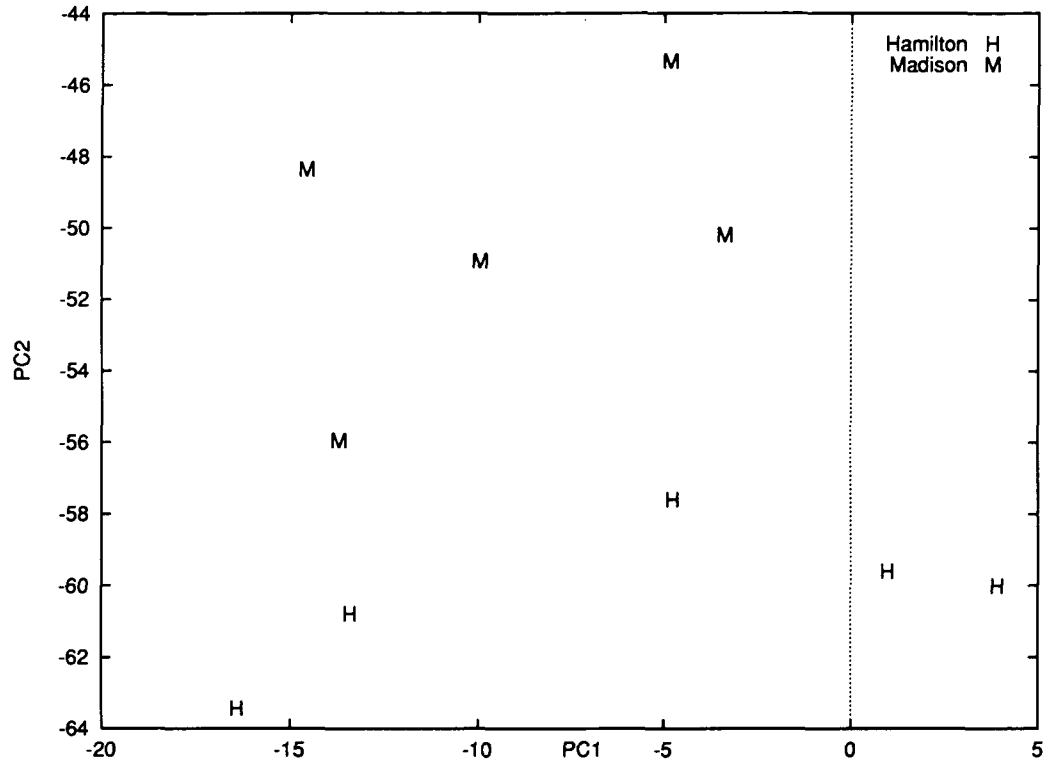

Fig. 7 Hamilton and Madison papers with thirty words.

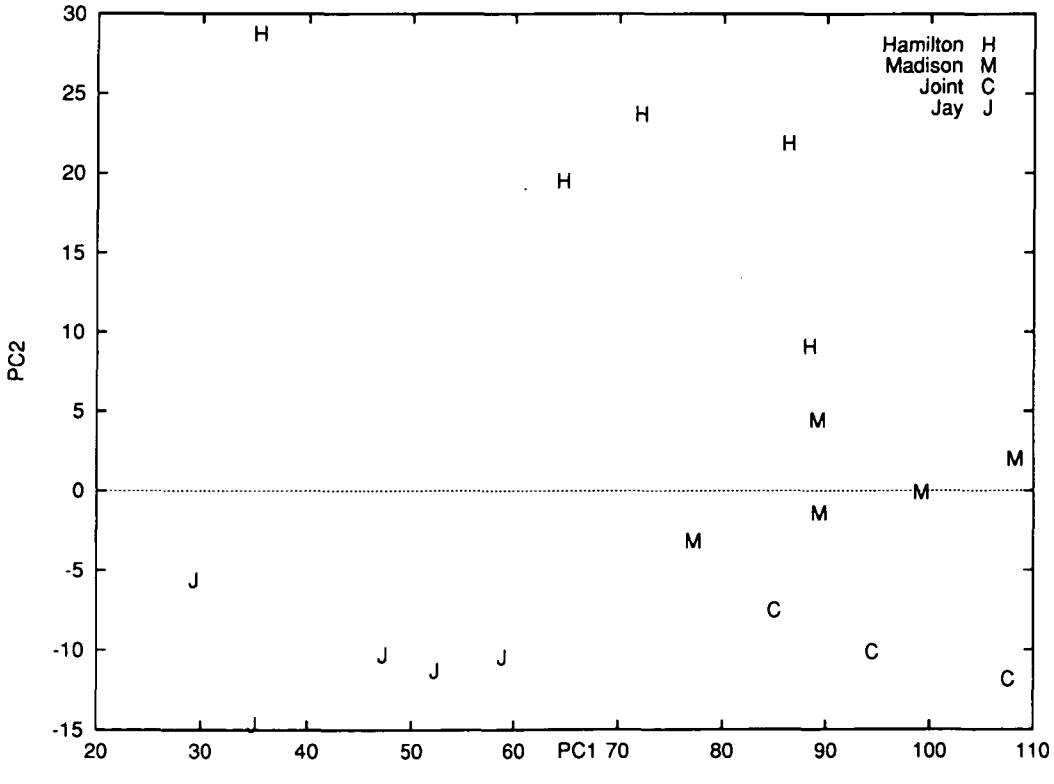

Fig. 8 Hamilton, Madison, Jay, and Joint papers with forty-nine words.

Burrows' contention that mutivariate word frequency analysis of large sets of common words is a stylometric technique which can discriminate between writers, even in such notoriously difficult cases as the *Federalist Papers*. Somewhat surprisingly, if we substitute the thirty marker words for the forty-nine-word set we obtain a plot in which the clusterings are less well defined.

The third set of results concerns the subset of the Hamilton, Madison and Disputed papers. Figure 9 shows the principal component plot for the forty-nine-word set, which now only accounts for 53.7% of the

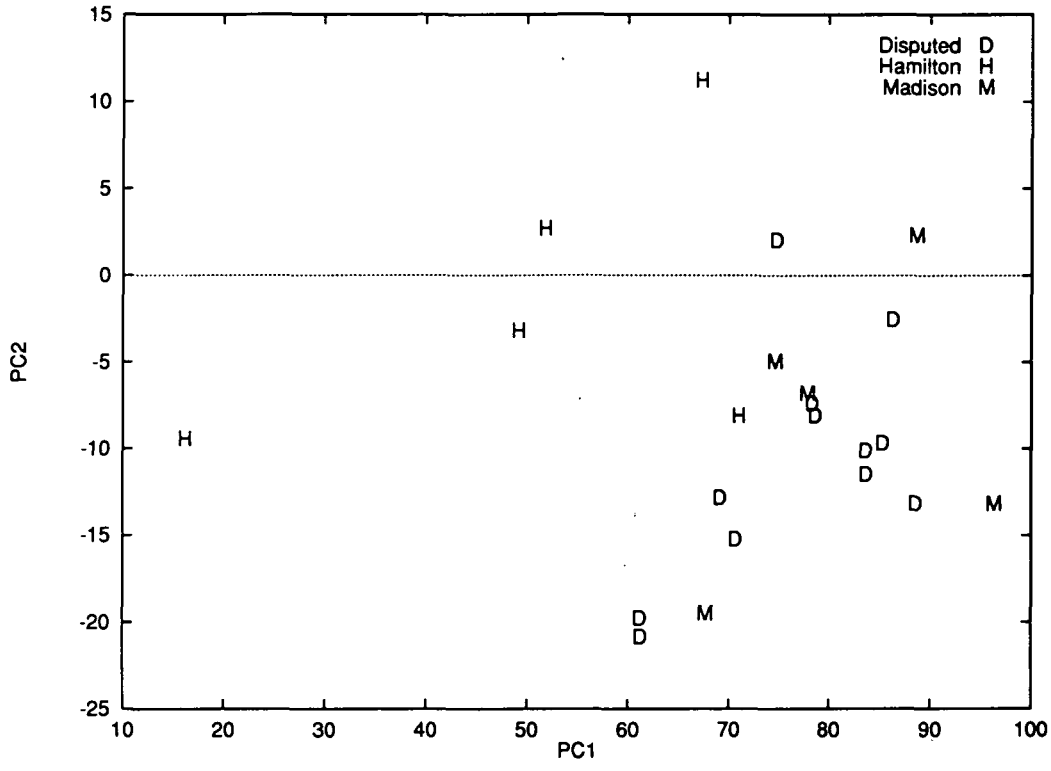

Fig. 9 Hamilton, Madison, and Disputed papers with forty-nine words.

total variation. Even so, the Disputed papers fall firmly in the Madison part of the plot along with Hamilton 70, a result in agreement with the conclusion of Mosteller and Wallace. The smaller data set of the thirty marker words produces a similar but less clear pattern, which though the first two components explain 77.1% of the variation. Once again we obtain somewhat tighter clustering with the larger set of words, even though they are not specifically chosen for their discriminatory power.

4. Finding Authorship Attribution Rules with a Genetic Algorithm

It is becoming recognized (Holmes, 1994) that computer-assisted authorship attribution can be seen as a pattern-recognition problem, in which the object of the exercise is to train some sort of classification program to distinguish between positive and negative examples of a particular author's work. It follows that techniques developed in the field of Machine Learning can usefully be applied to authorship attribution; and indeed studies within this paradigm have started to appear, e.g. Matthews and Merriam (1993), Kjell (1994), Merriam and Matthews (1994), Tweedie *et al.* (1994)—albeit only recently.

Most studies of this type have employed a 'connectionist' or 'neural' approach to machine learning. For example, Merriam and Matthews (1994) trained a Multi-layer Perceptron Network to distinguish between well-attested samples of writing by the Elizabethan dramatists William Shakespeare and Christopher Marlowe, and then applied the trained network to anonymous works and plays of dubious or mixed authorship, such as *Henry VI* Part 3, with illuminating results.

Accordingly, our third re-analysis of the Federalist Papers uses a machine-learning package but—in keeping with our intention of exploring some relatively unconventional techniques in stylometry—it does not involve a neural network. Instead we use a rule-finder package based on a 'genetic algorithm', the PC/BEAGLE system (Forsyth, 1987), to seek relational expressions characterizing the authorial styles of Hamilton and Madison.

4.1 *A Genetic Rule-Finder*

The BEAGLE system (Forsyth, 1981) can be seen as an early precursor of what would nowadays be called *genetic programming* (Koza, 1992). This software has been fully described elsewhere (Forsyth and Rada, 1986; Forsyth, 1989) so here we will only summarize its main features.

The package consists of six main modules, of which only two need concern us here: HERB (Heuristic Evolutionary Rule Breeder) which generates classification rules by a process modelled on the Darwinian idea of natural selection; and LEAF (Logical Evaluator And Forecaster) which applies rules produced by HERB to classify cases that may not have been seen before.

HERB uses an evolutionary search method that works with structures that encode Boolean relational expressions. It embodies the principles of the *evolution-strategy* (Rechenberg, 1973) and the *genetic algorithm* (Holland, 1975). HERB's learning algorithm is outlined below.

(i) Create an initial population of candidate rules at random (i.e. by making syntactically valid but, in general, semantically meaningless

combinations of operators, variables and constants).

(ii) Evaluate each rule on every training example and compute a fitness score based on the non-parametric correlation between the rule's truth value and the condition being predicted (with a penalty according to rule length, to encourage brevity).

(iii) Rank the rules in descending order of merit and remove the bottom half.

(iv) Replace 'dead' rules by crossing a pair of randomly selected survivors, thus assorting the recombining information from better rules.

(v) Mutate a small number of rules picked at random (excluding the best rule) and apply a tidying procedure to remove any redundancy thus introduced. Unless the termination conditions are satisfied, return to step (ii).

Each pass round this loop is called a 'generation' by analogy with the biological model. (For a good introduction to genetic algorithms, see Goldberg [1989].)

An example of a rule generated by HERB—taken from a preliminary run of the *Federalist* data where the system was given the task of finding Madison-indicating rules—is shown below.

$$((KIND < 0.00002)\&(TO < 35.205))$$

The rule above is true (indicating Madison's authorship) when the value of the variable *KIND* is less than 0.00002 and the value of the variable *TO* is less than 35.305; otherwise it is false (indicating non-Madisonian authorship). In this case the variables refer to the rate of usage, per 1000 words, of the words 'kind' and 'to'.

HERB requires data in standard 'flat-file' format, with each row representing a case or training instance and columns representing variables, as is usual for statistical packages. In order to present the program with data in a form it could use, an electronic version of the complete text of the *Federalist Papers* was obtained (from Project Gutenberg, at Illinois Benedictine College) and a Snobol4 program was written to compute frequencies of a user-supplied list of words in each paper. This pre-processing step transformed 1.2 megabytes of text into eighty-five rows of numbers, where each row represented an individual paper and each number in that row represented the usage rate (scales per 1000 words) of a particular word in that paper.

4.2 *The Calibration Process*

In all the runs reported below, HERB was run for 256 generations and set to produce three rules which were reduced to two by discarding the worst, i.e. that rule whose absence least degraded the classification performance on the training data. (This is a standard BEAGLE option.) Since BEAGLE rules yield binary results whereas this classification problem has three categories (Hamilton, Madison and Jay), both pro-Hamilton and a pro-Madison rule sets were generated in separate runs.

The data was first divided into sixty-nine undisputed papers for training and sixteen test cases. Of the training set, fifty-one were written by Hamilton, fourteen were by Madison and four were by Jay; of the test set, twelve were the disputed papers, three were Joint and one (number 64) was by Jay.

Previous experience with BEAGLE had shown that in cases where the proportion of positive to negative examples in the training data was unbalanced LEAF's probability estimates were often poorly calibrated. In this case the relative imbalance between authors (75% of the training instances being by Hamilton and only 20% by Madison) was felt to justify an alternative approach to probability estimation; so a simple form of cross-validation (Efron and Gong, 1983; Weiss and Kulikowski, 1991) was carried out in order to produce results that could be compared with the log-odds ratios quoted by Mosteller and Wallace.

This entailed a ten-fold replication of the following process.

First a division of the sixty-nine undisputed cases was made, with a random selection of (approximately) 80% being put into a training file and 20% held over as a validation subset. Then HERB was run for 256 generations on the training file. The best pair of rules was retained and applied to the unseen hold-over cases. A pair of rules each with a binary outcome can be in one of four states, Yes-Yes, Yes-No, No-Yes or No-No, but these four states were trichotomized as follows:

both rules Yes	position signal	(+)
both rules No	negative signal	(−)
rules disagree	mixed signal	(?)

On this basis, the results of each run on the unseen validation cases could be recorded in a 3 × 3 table cross-classified according to rule-decision (+, ?, −) and actual author (H, J, M). The numbers falling into each of nine cells were accumulated over ten runs.

The results using the thirty marker words especially selected by Mosteller and Wallace are given in section 4.3; the results using the forty-nine high-frequency function words, listed in Table 7 are given in Section 4.4.

4.3 *Results Using Mosteller and Wallace's Thirty Marker Words*

The results of this calibration process are tabulated in Tables 9 and 10, separately for pro-Hamilton and pro-Madison rules.

The point of this rather laborious procedure is that it gives us frequencies from which conditional probabilities for a Bayesian analysis, like that performed by Mosteller and Wallace (1984), can easily be extracted. Furthermore, these figures may be expected to be unbiased as they have been obtained from applying the rules to cases that were not used in the rule-creation process.

As an example, we can compute $P(M-|J)$, the probability of the pro-Madison rule-pair giving a negative signal on an unseen paper given that it was written by Jay, as

$$5/9 = 0.5556;$$

Table 9 Pro-Hamilton rule-calibration results

Pro-Hamilton Rule-Set:	Author = H	Author = J	Author = M
Both rules in favour ($H+$)	92	0	0
Rules disagree ($H?$)	11	4	6
Both rules against ($H-$)	1	5	29

Table 10 Pro-Madison rule-calibration results

Pro-Madison Rule-Set:	Author = H	Author = J	Author = M
Both rules in favour ($M+$)	1	1	12
Rules disagree ($M?$)	9	3	18
Both rules against ($M-$)	94	5	5

or we could calculate $P(H?|H)$, the probability of a pro-Hamilton rule-pair giving a mixed signal on an unseen paper by Hamilton as

$$11/104 = 0.1058.$$

Having obtained these calibration tables, the HERB program was run twice more (once to generate pro-Hamilton rules and once for pro-Madison rules), this time on all sixty-nine of the training cases. The rule-sets produced, again after 256 generations, were as shown in Table 11. (Interpretation of such rules will be postponed till Section 4.5; but meanwhile it should be said that the exclamation mark (!) is BEAGLE's logical negation operator and that the vertical bar (|) indicates an inclusive OR operation.)

Table 11 Rule-sets from 69 training cases

Pro-Hamilton Rules:
Rule 1: (*UPON* > *BOTH*)
Rule 2: ((*ON* − *THERE*) < 2.832)
Pro-Madison Rules:
Rule 1: ((*UPON* − *BOTH*) < *WHILST*)
Rule 2: !(*KIND*|((*BY* < 9.032) > *CONSEQUENTLY*))

Finally, both these rule-sets were applied to the sixteen papers not used either for training or calibration—the twelve Disputed papers, the three Joint papers and one paper by Jay which was left out of the training data (number 64).

The results on the disputed papers are summarized in Table 12.

The figures in the cells are the papers' serial numbers, thus, for instance, papers 50 and 55 were given mixed signals by both Hamilton and Madison rule-sets.

Table 12 Results of both rule-sets on Disputed papers

Pro-Madison / Pro-Hamilton	$M-$	$M?$	$M+$
$H+$			
$H?$		50,55	53
$H-$		54,56	49,51,52,57, 58,62,63

Overall, seven Disputed papers were simultaneously positively Madisonian (M+) and non-Hamiltonian (H−)—which is the most clear-cut and desirable result, given the prevailing opinion since 1964 that Madison was the true author of all twelve. Two were weakly Madisonian (M?) but positively non-Hamiltonian (H−); two were neither strongly Madisonian nor Hamiltonian (M? & H?) and one (number 53) was positively Madisonian (M+) as well as having a mixed signal regarding Hamilton's authorship (H?).

While only seven of the Disputed papers received the most conclusive possible attribution (M+ together with H−), it is notable that none of the twelve were more Hamiltonian than Madisonian while ten were more Madisonian than Hamiltonian.

As for the Joint papers, numbers 18 and 19 were both positively Madisonian (M+) and clearly non-Hamiltonian (H−) while number 20 was weakly Madisonian (M?) and strongly non-Hamiltonian (H−).

The single Jay paper tested, number 64, was given a positive signal by the pro-Madison rule-set (M+) and a negative signal by the pro-Hamilton rule-set (H−). On the face of it, this looks like a simple mistake, corresponding to an attribution to Madison (but definitely not Hamilton). This, however, is a case where using the calibration tables tells a rather different story than the bare categorization would suggest. The rules and tables were created with the discrimination between Hamilton and Madison primarily in mind, but conditional probabilities relating to Jay can be extracted to allow the same sort of inference as practised by Mosteller and Wallace in their 'Robust Bayesian Analysis' (as shown in Table 13).

Table 13 Conditional probabilities

Hamilton Rule-set:
$P(H-|Jay) = 5/9$ $P(H-|Not\ Jay) = 30/139$
Madison Rule-set:
$P(M+|Jay) = 1/9$ $P(M+|Not\ Jay) = 13/139$

Thus the odds-factor in favour of Jay, from the Hamilton rule-set is

$$\frac{(5/9)}{(30/139)} = 2.5741$$

to four places of decimals. This is the number by which one's prior odds in favor of Jay, as against either of the other two authors, would have to be multiplied to give posterior odds, taking this evidence into account. Using the Madison rule-set the same calculation is

$$\frac{(1/9)}{(13/139)} = 1.1880$$

to four decimal places. Perhaps surprisingly, the evidence, even from the Madison rule-set (with an outcome of M+), is somewhat in favour of Jay's authorship. The main reason for this appearance of paradox is that labelling M+ 'positively Madisonian' is just a convenient shorthand: the import of the evidence provided by an outcome of M+ from the pro-Madison rule-set

is most fully contained in the relevant calibration table.

Since the odds-factors derived from the two rule-sets do not agree, the question arises of how to combine them. If Hamilton and Madison rule-sets could be treated as independent items of evidence, we could simply multiply them, giving a resultant factor of 3.058 in Jay's favour. However, this assumption of independence is hardly tenable when that the two rule-sets share two varibles, namely UPON and BOTH. Alternatively, if joint outcomes of both rule-sets had been gathered and tabulated (as a 9 × 3 table) during the calibration process then a combined odds-factor could have been calculated—but that really would have been a tedious task, giving the present facilities of the BEAGLE software.

A reasonable compromise in the circumstances is to form a composite odds-factor by, in effect, merging the rows selected by the rule outcomes in the two separate calibration tables. This gives a JOF (Jay Odds Factor) of

$$\frac{(6/18)}{(43/278)} = 2.155$$

for paper number 64, making our posterior odds on Jay's authorship more than twice as large as our prior odds, whatever they might have been.

More pertinently, applying the same reasoning to the disputed papers yields the results listed in Table 14.

Table 14 Madisonian odds-factors for Disputed papers

Papers	Factor from Madison rules	Factor from Hamilton rules	Composite Factor
50,55	5.9429	1.6208	3.5657
53	35.6571	1.6208	4.4571
54,56	5.9429	86.1714	13.9657
49,51,52, 57,58,62,63	35.6571	86.1714	60.9143

Here the main cell entries are MOFs (Madison Odds Factors) provided by the two different rule-sets. The final column, labelled 'composite', is obtained by using the same method as described above for paper number 64 by Jay. It is an inherently conservative estimate, corresponding to taking an average odds-factor weighted according to the number of cases (of both categories) falling into the relevant rows of the calibration table.

It will be seen that all twelve papers have odds factors greater than 1, meaning that the posterior odds in favour of Madison as opposed to Hamilton are greater than the odds before this evidence was obtained. For seven of these disputed papers, the majority, the MOF is between 35.6571 and 86.1714. If we accept, for the sake of argument, the composite odds factors (which, if biased at all are likely to be biased towards understatement in whichever direction they point) then the results range from odds factors of 3.57 to 60.91 in favour of Madison.

In the analysis by Mosteller and Wallace which this study most closely resembles, their 'Robust Bayesian Analysis', the odds factor of the most clearly Madisonian paper, number 54, were 169,396—a far more conclusive attribution than ours. However the lowest Madisonian odds-factor of 0.1003 was given to paper number 55 (also one of our 'problem cases'), representing a factor of 10 to 1 in Hamilton's favour.

When it is borne in mind that Mosteller and Wallace used thirty-one marker words in that particular analysis (after screening) while BEAGLE whittled a set of thirty candidate words down to eight (four in the pro-Hamilton rules, six in the pro-Madison rules with two appearing in both) then it may fairly be claimed that the present method does have something to offer in investigations of this kind.

4.4 *Results Using Forty-nine High-Frequency Function Words*

Essentially the same analysis was repeated with one difference: instead of using Mosteller and Wallace's thirty selected marker words as variables, forty-nine high-frequency function words (see Section 3) were used. The idea was to find out whether good results could be obtained without special effort at word-selection, emulating Burrows (1992). Calibration tables are given in Tables 15 and 16.

Table 15 Calibration table for pro-Hamilton rule-set

Pro-Hamilton Rule-Set:	Author = H	Author = J	Author = M
Both rules in-favour ($H+$)	73	1	4
Rules disagree ($H?$)	30	3	16
Both rules against ($H-$)	1	0	12

Table 16 Calibration table for pro-Madison rule-set

Pro-Hamilton Rule-Set:	Author = H	Author = J	Author = M
Both rules in favour ($M+$)	0	0	7
Rules disagree ($M?$)	11	2	15
Both rules against ($M-$)	93	2	10

The actual rules produced were as shown in Table 17.

Table 17 Actual rules produced using forty-nine function words

Pro-Hamilton Rules:
Rule 1: ((*ON* < = 6.9517)&((56.5475 − *OF*) < 2.0729))
Rule 2: ((*BY* − *TO*) < = −25.7484)
Pro-Madison Rules:
Rule 1: (*ON* > = 6.9202)
Rule 2: !(((*IN* − *AND*) > = −6.6181)|((*THEIR* < *WOULD*) = *WAS*)

Applied to the Disputed papers the joint outcomes from both rule sets were as shown in Tables 18 and 19.

Here the majority of assignments are much less confident than in Section 4.3; and we also have two definite mistakes, on paper number 55 (which Mosteller and Wallace found the least Madisonian of the twelve) and number 63 (the last that Madison contributed to the series). It would seem that HERB did benefit in the first experiment from having specially selected marker

Table 18 Results of both rule-sets with forty-nine function words

Pro-Madison / Pro-Hamilton	$M-$	$M?$	$M+$
$H+$		55	
$H?$	63	52,53	
$H-$		49,50,51, 54,57,58,62	56

Table 19 Madisonian odds-factors on Disputed papers using forty-nine function words

Papers	Factor from Hamilton rules	Factor from Madison rules	Composite Factor
55	0.1781	4.4318	0.7351
63	1.7333	0.3495	0.6870
52,53	1.7333	4.4318	2.4573
49,50,51, 54,57,58,62	39.0	4.4318	7.3125
56	39.0	∞	61.75

words to chose from. In particular the loss of four of the five best markers ('there', 'upon', 'while', and 'whilst') does appear to have hampered the learning process.

In addition, the rules are less succinct and thus harder to understand than those produced using the thirty selected Mosteller-and-Wallace markers.

Nevertheless a crude 'success rate' of 10/12 without any pre-selection of variables in what is acknowledged to be a difficult attribution problem is by no means disgraceful.

4.5 *The Role of Rules*

Almost as important as accuracy is the form that the knowledge required by a classifier takes. In the literature of machine learning and expert systems, knowledge representation is a central concern (e.g. Parsaye and Chignell, 1988; Ignizio, 1991) but in authorship studies it has taken a back seat. The very fact of being able to generate compact and (reasonably) intelligible rules from examples is in itself a new departure in stylometry.

Stylometers have used a variety of indicators over the years, including marker words, measures of vocabulary distribution, collocations and proportionate pairs. A Boolean expression like

$$((UPON\text{–}BOTH) < WHILST)$$

is none of these. Nevertheless it is a very economical summarization of some of the more characteristic of Madison's verbal habits. It says simply that the difference between the rate per 1000 words of 'upon' and 'both' is typically less than the rate of 'whilst' in a passage written by Madison. The computer has discovered a concise three-term relationship between word rates, whose interpretation can be illustrated if we apply the mean rates per 1000 words for the two main Federalist authors (from Mosteller and Wallace, 1984), to the formula in Table 20 and work out its truth value.

The ability to find possibly novel combinations of stylistic indicators such as that above is, we contend, a useful addition to the stylometer's toolkit. Thus programs like BEAGLE, and especially more advanced successors to it, will have a part to play in the development of this field.

Table 20 Pro-Madison rule evaluated with mean rates

Hamilton mean rates:
3.35 − 0.47 < 0 ⇒ False
Madison mean rates:
0.14 − 1.08 < 0.48 ⇒ True

Some advantages of rule-induction systems like BEAGLE are:

- a relatively large number of potential indicator variables can be reduced to a much smaller subset which are effective in combination with each other;
- logical rules can be communicated and criticized in ways that neural-network weight-matrices and numeric discriminant functions cannot (hence their popularity in expert systems);
- there is nothing to prevent indicators such as collocations, measures of vocabulary richness and so forth being given to a rule induction program as variables instead of or as well as individual word frequencies, even though we did not do that here;
- rules give some idea of the relationships existing between linguistic variables, in theory permitting the discovery of diagnostic combinations of variables which in themselves are only weakly indicative.

The weaknesses of such systems (with special reference to BEAGLE) include the following:

- without a calibration table or something equivalent, misleading attributions are quite likely;
- like most such software, BEAGLE does not make the creation of a calibration table an easy task (though the system's author is now well appraised of this deficiency, which is curable given time!);
- BEAGLE is not well integrated into a coherent inferential framework, such as Bayesian inference;
- since rules do not specify their context, there is a danger of applying them outside their scope and being misled.

This last point is by no means unique to rule-induction systems: it applies to almost any conceivable classification technique. For example, Matthews and Merriam (1993) trained a neural network to produce a 'Shakespeare Characteristic Measure' (SCM) in distinguishing between works by Shakespeare and by Fletcher; but when they wanted to distinguish Shakespeare from Marlowe (Merriam and Matthews, 1994) they created a completely different SCM. In other words, the SCM developed versus Fletcher might well be useless as an SCM versus Marlowe; just as a Hamilton (versus Madison) rule-set might prove useless for distinguishing Hamilton from, for instance, Jefferson. Although this point seems quite obvious in cold print, the issue of the range of applicability of automatically derived classifiers is a practical problem and is likely to become more serious as inductive systems gain wider acceptance.

The way forward, we believe, is to attempt to minimize or counteract these drawbacks whilst preserving

some of the undoubted advantages of the rule-induction approach—as we hope to show in future articles.

Conclusion

Although the *Federalist* problem is acknowledged to be a severe choice on which to test attribution techniques, all three of our approaches have produced encouraging results. The data-preparation underpinning both the vocabulary richness and word frequency analysis techniques necessitates working with samples of the eighty-five papers involved and our findings must be qualified in this respect. Nevertheless, the multivariate analyses of the vocabulary richness measures have hinted at a group structure amongst the candidate authors with the Disputed papers showing a 'Madison' tendency, in agreement with the findings of Mosteller and Wallace. The disciminant analysis was particularly impressive, suggesting that, in a collective sense, vocabulary richness variables provide a good set of discriminators.

Word frequency analysis using the forty-nine high-frequency function words produces quite clear clustering with the Disputed papers, once again, falling firmly on the 'Madison' side. Remarkably this clustering is less clear with the thirty Mosteller and Wallace marker words even though they were originally chosen for their discriminatory power. Perhaps large ($\geqslant$ 50) sets of common words must be employed in order to obtain meaningful results with this technique, a finding which would concur with Burrows' work in this area.

The genetic rule-finder method was able to use all eighty-five papers, sixty-nine for training and sixteen as test cases. Given the prevailing opinion that Madison is the true author of all twelve Disputed papers, the rule-sets generated and then employed on the thirty marker words performed 'better' than the rule-sets using the forty-nine-word set, a reverse finding to that above. In the first case, none of the twelve were more Hamiltonian than Madisonian, whilst in the second case two papers (numbers 55 and 63) fell into that category. Considering that only eight marker words were ultimately employed from the thirty words available in the first of these cases, the result was most encouraging.

Machine-learning methods in general, and genetic-based machine learning in particular, have been under-utilized in stylometric studies hitherto: our research suggests that the prospects for their successful application in future look good. We intend to continue this research programme by feeding other indicators of style into the rule-induction program, not just individual word frequencies. This would provide a fascinating scenario where vocabulary richness measures were employed in a genetic algorithm to seek characterizing relational expressions and would blend two of the approaches to authorship attribution covered in this paper.

Acknowledgements

Our grateful thanks go to Fiona Tweedie and Lisa Robertson of the Department of Mathematical Sciences at the University of the West of England, Bristol, for their help in the preparation of this paper. We also wish to thank Professor Robert Valenza of Claremont McKenna College, California, and Project Gutenberg at Illinois Benedictine College for electronic versions of the *Federalist Papers*.

References

Bayes, T. (1763). An Essay Towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society*, 53: 370–418.

Brunet, E. (1978). *Vocabulaire de Jean Giraudoux: Structure et Évolution*. Slatkine, Genève.

Burrows, J. F. (1992). Not Unless You Ask Nicely: The Interpretative Nexus between Analysis and Information. *Literary and Linguistic Computing*, 7(2): 91–109.

Damerau, F. J. (1975). The use of Function Word Frequencies as Indicators of Style. *Computers and the Humanities*, 9: 271–280.

Efron, B. and Gong, G. (1983). A Leisurely Look at the Bootstrap, the Jack-knife and Cross-validation. *The American Statistician*, 37(1): 363–38.

Forsyth, R. S. (1981). BEAGLE, a Darwinian Approach to Pattern Recognition. Kybernetes, 10: 159–166.

—— (1987). *PC/BEAGLE User Guide*. Pathway Research, Nottingham.

—— (1989). *Machine Learning: Principles and Techniques*. Chapman and Hall, London.

—— and Rada, R. (1986). *Machine Learning: Applications in Expert Systems and Information Retrieval*. Ellis Horwood, Chichester.

Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison Wesley, Reading, MA.

Holland, J. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor.

Holmes, D. I. (1991). Vocabulary Richness and the Prophetic Voice. *Literary and Linguistic Computing*, 6(4): 259–268.

—— (1992). A Stylometric Analysis of Mormon Scripture and Related Texts. *Journal of the Royal Statistical Society Series A*, 155(1): 91–120.

—— (1994). Authorship Attribution. *Computers and the Humanities*, 28(2): 87–106.

Ignizio, J. P. (1991). *Introduction to Expert Systems*. McGraw Hill, New York.

Kjell, B. (1994). Authorship Determination using Letter Pair Frequency Features with Neural Network Classifiers. *Literary and Linguistic Computing*, 9(2): 119–124.

Koza, J. (1992). *Genetic Programming*. MIT Press, Cambridge, MA.

Matthews, R. and Merriam, T. (1993). Neural Computation in Stylometry I: An Application to the Works of Shakespeare and Fletcher. *Literary and Linguistic Computing*, 8(4): 203–209.

McColly, W. B. and Weier, D. (1983). Literary Attribution and Likelihood Ratio Tests—the Case of the Middle English Pearl-poems. *Computers and the Humanities*, 17: 65–75.

Merriam, T. V. N. (1989). An Experiment with the Federalist Papers. *Computers and the Humanities*, 23: 251–254.

—— and Matthews, R. (1994). Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Marlowe. *Literary and Linguistic Computing*, 9(1): 1–6.

Morton, A. Q. (1978). *Literary Detection*. Bowker,

Mosteller, F. and Wallace, D. L. (1984). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Addison-Wesley, Reading, MA.

Parsaye, K. and Chignell, M. (1988). *Expert Systems for Experts*. Wiley, New York.

Rechenberg, I. (1973). *Evolutionsstrategie — Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Frommann-Holzboog, Stuttgart.

Sichel, H. S. (1975). On a Distribution Law for Word Frequencies. *Journal of the American Statistical Association*, 70: 542–547.

—— (1986). Word Frequency Distributions and Type-token Characteristics. *The Mathematical Scientist*, 11: 45–72.

Tweedie, F. J., Singh, S. and Holmes, D. I. (1994). Neural Network Applications in Stylometry: The Federalist Papers. In Monaghan, A. I. C. (ed), *CSNLP*. Natural Language Group, Dublin City University.

Weiss, S. M. and Kulikowski, C. A. (1991). *Computer Systems that Learn*. Morgan Kaufmann, San Mateo, CA.

Wills, G. (1982). *The Federalist Papers of Alexander Hamilton, James Madison and John Jay*. Bantam Books, Toronto.