Computers and the Humanities **38:** 15–36, 2004.

Detecting Collaborations in Text

Comparing the Authors' Rhetorical Language Choices in The Federalist Papers

JEFF COLLINS[1], DAVID KAUFER[1], PANTELIS VLACHOS[1], BRIAN BUTLER[2] and SUGURU ISHIZAKI[3]
[1] *Carnegie Mellon University, USA*
E-mail: jeff.collins@acm.org
[2] *University of Pittsburgh, USA*
[3] *Pittsburgh, Pennsylvania, USA*

Abstract. In author attribution studies function words or lexical measures are often used to differentiate the authors' textual fingerprints. These studies can be thought of as quantifying the texts, representing the text with measured variables that stand for specific textual features. The resulting quantifications, while proven useful for statistically differentiating among the texts, bear no resemblance to the understanding a human reader – even an astute one – would develop while reading the texts. In this paper we present an attribution study that, instead, characterizes the texts according to the representational language choices of the authors, similar to a way we believe close human readers come to know a text and distinguish its rhetorical purpose. From our automated quantification of *The Federalist* papers, it is clear why human readers find it impossible to distinguish the authorship of the disputed papers. Our findings suggest that changes occur in the processes of rhetorical invention when undertaken in collaborative situations. This points to a need to re-evaluate the premise of autonomous authorship that has informed attribution studies of *The Federalist* case.

Key words: authorship attribution, collaboration, federalist papers, statistics

1. Introduction

Opening their book on the disputed authorship of *The Federalist* papers, Frederick Mosteller and David Wallace write, "When two statisticians, both flanks unguarded, blunder into an historical and literary controversy, merciless slaughter is imminent" (1964, p. 1). Of course no such slaughter occurred and their galvanizing findings have been a benchmark for statisticians and language theorists interested in testing attribution methodologies and automated systems of language analysis.

Like others, we initially approached *The Federalist* case as a benchmark: we've been developing a catalog to classify text strings according to their primary representational purpose, a set of language features we believe close human readers use to grasp rhetorical intentions as they read.

The language theory behind our catalog is drawn from a long history of rhetorical interest in patterns of language effect. The rhetorician I.A. Richards (1991), writing in the mid-1930s, suggested people read texts with an implicit faith, an assumption, that the author intended to convey a meaning and made language choices consistent with conveying it. Drawing broadly from rhetorical tradition and practice, two of the authors built a representational theory of composition that describes the rhetorical patterns within texts (see Appendix A). Their book-length exploration identified the limited variety of patterned, interactive experiences language makes available to readers (Kaufer and Butler, 2000).

We have observed that experienced writers control these "representational" effects, manipulating their writing here and there to achieve subtle rhetorical impressions that are cumulative: no one language choice necessarily makes a strong impression, but cumulatively the choices lead to particular impressions for readers. Teaching student writers such subtle control of language is difficult. Historically, this control has been attained through years of reading and writing practice in the school setting and beyond it. Our language theory and software are designed to try to speed the learning process by helping our writing students understand these cumulative language effects more explicitly, giving them instantaneous, consistent feedback on their use of representational effects in their writing. This theory of representational language effects is the basis of our analysis of *The Federalist* papers.

The authorship problem presented by the dispute over *The Federalist* papers is an appropriate case study for an exploration of our language theory because the papers were written to achieve a specific rhetorical plan. Douglass Adair (1944) describes the great lengths to which *The Federalist* authors, Alexander Hamilton and James Madison, went to hide their own voices behind the collective pseudonym "Publius," enlivening this persona with all the persuasive reasoning and oratorical gusto needed to convince the legislature of New York to ratify the recently drafted U.S. Constitution. The authors left even their contemporary readers stumped about authorship: George Washington asked an aide, "Pray, if it is not a secret, who is the author or authors of Publius?" (qtd in Adair, p. 237n6). Since the controversy over the authorship of 12 disputed papers arose, historians and other careful human readers have been unable to discern which of the authors wrote the disputed texts (Adair, 1944, p. 247 and following). Therefore, using *The Federalist* as a test case would provide an indication of the sensitivity of our language theory and computer software as tools for characterizing the representational language choices of authors attempting to achieve specific rhetorical plans.

Several authors (Holmes, 1994; Holmes and Forsyth, 1995; Martindale and McKenzie, 1995; Rudman, 1998) provide overviews of author attribution methods for which we are grateful. In this study, we compare the frequencies with which the two authors of *The Federalist* papers use representational language strings. The results are intriguing because they indicate significant differences between the authors in their known papers ($p < 0.05$) while finding overlapping usages

within the disputed papers. Based upon our findings, we argue that the nature of the collaboration between the two authors simply does not allow for clear separation, a position consistent with that of historians (e.g. Carey, 1984; Kesler, 1999).

In typical attribution studies investigators employ measures such as function word frequency, lexical statistics, or vocabulary richness to decide upon an author. The quantifications serve as a sort of smoking gun that allows the investigators to point with some calculated degree of certainty to the culprit responsible for the text. Such attribution studies are premised on the commonsense assumption that writing is an individual act (cf. Mallon, 2001; Woodmansee, 1994; Rudman, 2000, pp. 164–166). Even in collaborative projects such as *The Federalist*, attribution studies seem to assume that one or another of the collaborators in each text must have been the dominant author and can, therefore, be exposed through attribution methods, if only the methods are good enough.

Our findings more closely resemble and would seem to support contemporary theoretical investigations of collaboratively composed texts (Ede and Lunsford, 1990; Stillinger, 1991; Irizarry, 1993) and notions of rhetorical invention as social acts (LeFevre, 1987; Kaufer and Carley, 1993) in which case the question underlying attribution studies of *The Federalist* should not be "Which author should get credit for a paper?" – with its accompanying implication of exclusive, atomistic authorship – but "Can the attribution method detect collaboration in the products of the effort?" As we shall show, our analysis produces a nuanced picture of the collaboration in *The Federalist* papers, one that not only accounts for current understandings of the collaborative nature of rhetorical invention, but one that is also more congruous with the documented history of *The Federalist* project.

2. Methods

2.1. Language theory and software

In this section, we describe how the rhetorical theory of representational language usage operates. This method of characterizing language choice is drawn from rhetoricians' long-standing interest in the patterns of language that prime interactive experiences for an audience (Kaufer and Butler, 1996, 2000).

Designing a tagging catalog to distinguish the usages suggested by this representational theory is challenging because an individual word might be part of many different effects. Consider the word "smear." In a sentence like "John smeared his opponent," the word primes a markedly different effect than when used in a sentence like "John smeared butter on his toast." The first "smear" indicates a negative affect while the second indicates a physical motion. An information system that attempts to tag effects using single words would miss the rhetorical difference.

Some parsers, looking at rhetorical or grammatical structure (e.g. Marcu, 2000), have the flexibility needed to parse the sentences, but provide a focus on categories that are blind to representational effect. We designed our catalog specifically to

identify the language strings authors use to prime distinct representational effects for readers (listed in Appendix A). For example, word strings such as those given below would appear in the catalog, differentiating two of the representational effects involving the word *smeared*:

Motion
smeared butter
smeared margarine
smeared tanning oil
smeared jam
smeared [article] butter
smeared [article] margarine
etc.

Negative Affect
smeared [pronoun] opponent
smeared [pronoun] enemy
smeared [pronoun] opposition
smeared [article] opponent
smeared [article] enemy
smeared [article] opposition
etc.

Over the past decade, we have compiled over 250 million of such language strings into our catalog, classifying each string into 1 of the 18 non-overlapping categories of representational effect (which serve as the response variables in this study). Using our tagging and visualization software, DocuScope (see Collins, 2001), for information), provides writers with a useful estimate of the interactive experiences their texts will prime for readers.

In highlighting the strings responsible for these 18 representational effects, the software makes available a portion of the information we believe readers attempt to gain through close readings of texts. We are by no means suggesting our theory makes available all the information from a close reading: it captures little of style, for example, and it does not attempt to interpret language for deep semantics. Nor can it match the rhetorical structures that depend on discontinuous elements across sentences and paragraphs. The strings captured by our theory represent what a myopic reader of English with an impoverished understanding of structure, context and culture could find.

Although this "reading" is challenged in these ways, the software finds and tags these strings more consistently than could most human readers. Our software automatically applies our representational catalog to collections of texts, tagging each text and indicating the quantity of each of the 18 variables found in each text (given as a percentage). This enables careful assessment of an authors' language choices in a way similar to the use of other dictionary-based tagging programs and parsers like Wordnet or Diction (Stone, 2000; Fellbaum, 1998; Scolari, 2000). The quantifications that result from tagging texts allow for statistical comparisons of the tagged textual features. An example tagged sentence and a text quantification is provided in Appendix B.

We used the electronic version of *The Federalist* papers available from Project Gutenberg (Turner, 1992). When applied to the papers, our software tagged approximately 27% of each paper with no significant difference in the amount of text tagged between the authors ($p = 0.21$). The software saves these quantifica-

tions as a comma-delimited text file that can be analyzed using a statistical package or spreadsheet program. The quantification of *The Federalist* papers that underlies our analysis is available from the datasets archive of Carnegie Mellon's StatLib (Vlachos, 2001).

2.2. Statistical Methods

Unlike Mosteller and Wallace (1964), we confined our analysis to *The Federalist* texts, despite the smaller sample size this provides for Madison. We eschewed other writing samples because texts outside *The Federalist* would have been written for different rhetorical purposes, confounding the results of our exploration. We also excluded consideration of the three so-called "joint" papers because considering such a small group of texts which were written so closely together and under a different condition of authorship would introduce several potentially-confounding uncontrolled variables. Likewise, we did not analyze the contributions of the third collaborator in the project, John Jay, because he is not involved in the authorship dispute and wrote only five papers. We, therefore, examined three groups of papers: the undisputed Hamilton papers; the undisputed Madison papers; and the disputed papers.

We relied on three statistical methods for making comparisons of *The Federalist* texts. We used multivariate and univariate analyses of variance to compare the groups of papers. We also used a procedure known as "discriminant analysis" (Kachigan, 1991, pp. 216–235) to describe the relationships among the variables involved in distinguishing one author's texts from the other's. This procedure resulted in a mathematical formula (given in Appendix C) that provides a score for each individual paper. We used these scores in a third procedure, called "the bootstrap" (Efron, 1979; Efron *et al.*, 2001), to enable accurate comparisons between the individual paper scores and large, normally-distributed samples of scores. The bootstrap provides for an estimate of the confidence that a particular paper's score falls among a particular group. We describe each of these procedures in more detail in the following sections.

3. Statistical Results

All of our statistical analyses are based upon the papers' scores on the 18 representational dimensions described in Appendix A. These scores (summarized in Table I) serve as the response variables in the following procedures.

3.1. Analyses of Variance

To compare the groups of papers on all of the response variables simultaneously, one-way multivariate analysis of variance (MANOVA) was carried out. MANOVA indicates whether or not the three groups of texts (Hamilton, Madison, disputed)

Table I. Variable means and standard deviations for the three groups of papers

Variable	Mean (standard deviation)		
	Hamilton, $n = 51$	Madison, $n = 14$	Disputed, $n = 12$
First Person	0.168 (0.17)	0.103 (0.08)	0.140 (0.07)
Inner Thinking	2.512 (0.50)	2.412 (0.27)	2.614 (0.42)
Think Positive	1.741 (0.48)	1.655 (0.51)	1.468 (0.38)
Think Negative*	2.146 (0.69)	1.844 (0.84)	1.535 (0.57)
Thinking Ahead*	1.339 (0.37)	1.073 (0.40)	1.103 (0.38)
Thinking Back	0.481 (0.20)	0.474 (0.19)	0.363 (0.17)
Word Picture	1.410 (0.40)	1.314 (0.44)	1.473 (0.62)
Space Interval	0.343 (0.15)	0.357 (0.16)	0.307 (0.14)
Motion	0.164 (0.11)	0.124 (0.06)	0.142 (0.11)
Past Events**	1.122 (0.25)	1.578 (0.29)	1.495 (0.35)
Shifting Events	0.336 (0.14)	0.369 (0.10)	0.412 (0.19)
Time Interval	0.698 (0.26)	0.554 (0.14)	0.645 (0.30)
Cue Com Knwldge	1.842 (0.42)	1.704 (0.39)	1.728 (0.41)
Cue Prior Text	4.849 (0.69)	4.776 (0.79)	4.693 (0.49)
Cue Reader	0.681 (0.45)	0.565 (0.35)	0.553 (0.34)
Cue Notifier	6.325 (0.73)	6.826 (0.71)	6.467 (0.62)
Cue Movement	0.012 (0.02)	0.012 (0.02)	0.018 (0.03)
Cue Reasoning	0.739 (0.24)	0.824 (0.16)	0.765 (0.21)

$df = (2, 74)$; * = $p < 0.05$; ** = $p < 0.001$.

differ significantly on the variables. We chose the most conventional statistic, the Wilks' Lambda test Wilks, 1932), which yielded $F(36, 114) = 2.110$, $p = 0.002$. The significant result on this test does not mean that each group differs significantly on each of the 18 variables. To test for this possibility a univariate analysis of variance (ANOVA) was performed for each response variable. Three of the 18 variables were found to be significant, shown in Table II.

When sample sizes are nearly equal, analyses of variance are not overly sensitive to inequalities of variance. However, when one sample is much larger than another (as in our case) the calculated F statistic may be dominated by the variances of the larger sample (Hamilton). We evaluated this possibility by randomly sampling 14 Hamilton papers 1,000 times and calculating the Wilks' Lambda score each time. This Monte Carlo procedure provided a distribution of Wilks' Lambda scores and the variability for a balanced design. The procedure resulted in a mean Lamda that was significant marginally ($p = 0.08$).

For each of the three significant variables (*Think Negative*, *Thinking Ahead*, and *Past Events*) we used Tukey's method (Tukey, 1949) to make pairwise comparisons between the groups of papers. This comparison method provides an indication

Table II. F-ratios, p-values, and Tukey's Method results for the significant variables

Variable	$F(2, 74)$	p	Tukey's method[a]
Think Negative	4.10	0.020	
Thinking Ahead	3.91	0.024	
Past Events	20.40	< 0.001	H–M, H–D

[a]Differences between groups significant at $p < 0.05$ or better.

of whether a pair of groups – e.g., Hamilton and Madison – differ significantly by constructing confidence intervals for all the pairwise differences between the variable means. In other words, this test tells us if each pair of groups differs on the variable being considered.

Tukey's method showed significant differences (at $\alpha < 0.05$ or better) between the Hamilton and Madison groups and also between the Hamilton and the disputed groups on *Past Events*, but not on the other variables. The method also indicated no significant difference exists between the disputed papers and the Madison papers.

From our results it is safe to say that as a group the disputed papers contain a frequency of *Past Events* language strings that are more similar to Madison's texts than to Hamilton's. This finding would tend to support the view engendered by Mosteller and Wallace's (1964) study that Madison is the likely author of the disputed papers. However, our results to this point cannot be reliably used to estimate the probability of authorship of each disputed paper.

3.2. Discriminant Analysis

Discriminant analysis (Kachigan, 1991, pp. 216–235) is a procedure for defining a boundary between groups by identifying the relationships between quantitative response variables and a criterion variable that may be qualitative in nature (in our case, the author of the paper). We used discriminant analysis to find a distinct boundary separating the undisputed papers of Hamilton and Madison. Defining such a boundary would allow us to attempt to classify the authorship of each of the disputed papers.

We found several of the 18 response variables to be co-linear, which would cause unsatisfactory results in attempting to develop a discriminant function. Therefore, we used a stepwise procedure (Seber, 1984) to select a subset of the variables that were not co-linear and that would likely be the most useful for discriminating between the two authors. Using a stepwise procedure, the response variables are subjected to a sequence of analysis of covariance tests. This results in the elimination of variables that are redundant with other variables. In our study, 7 of the 18 variables were thus eliminated. The remaining 11 variables (listed in Appendix C) were found to provide satisfactory convergence so that we could proceed in developing the discriminant function.

Table III. Means, standard deviations, and confidence intervals of the logistic function scores

Group	*n*	Mean	St. dev.	95% C.I.	
				Lower	Upper
Hamilton	51	–7.99	5.58	–18.93	2.95
Madison	14	5.32	7.54	–9.46	20.10
Disputed	12	1.51	8.54	–15.23	18.25

Table IV. Posterior probabilities of the logistic function using Hamilton and Madison undisputed papers as validation sample

Actual group	Predicted group		Correct	Misclassified papers
	Hamilton	Madison		
Hamilton	50	1	98.0%	#23
Madison	3	11	78.6%	#14, #41, #46

Since our criterion variable (*author*) is binary – either Hamilton or Madison – we used logistic regression for this procedure (Hosmer and Lemeshow, 2000). Logistic regression is a technique for developing a mathematical function that discriminates between two groups of samples. The technique investigates the relationship between the response variables and a criterion variable. Unlike more common techniques of linear regression, logistic regression avoids generating erroneous probability predictions that might result from having a binary criterion variable or from having response variables that are not all normally distributed (Fienberg, 1979; O'Hara *et al.*, 1982).

In our case, we used logistic regression with the two authors' undisputed texts to develop a mathematical function that would discriminate between the papers by Hamilton and Madison. The function and an example of its use may be found in Appendix C. The result of applying this discriminant function to each of *The Federalist* papers is summarized in Table III.

We tested the efficacy of this discriminant function by applying the function to the two groups of undisputed papers and calculating the posterior probabilities of the function making the correct discrimination. The results are shown in Table IV. The likelihood ratio test for the function yields $\chi^2 = 50.173$ ($df = 11$, $p < 0.001$). It would have been better to validate the discriminant function with a sample of papers that were not used in its development (see Kachigan, 1991, p. 232); unfortunately the limited number of papers available does not permit such a cross-validation. Therefore, the true discriminatory power of the function would be

Table V. Means, standard deviations, and confidence intervals of the logistic function scores for the bootstrap samples ($B = 1,000$ per group)

Group	Mean	St. dev.	95% C.I.	
			Lower	Upper
Hamilton	–7.78	1.30	–10.32	–5.24
Madison	4.02	1.73	0.64	7.40
Disputed	0.49	2.95	–5.28	6.27

somewhat less accurate than is indicated in Table IV. Nonetheless, the discriminant function performed adequately to proceed.

We calculated the logistic function scores for each of the disputed papers (provided in column 2 of Table VII). The results of these calculations indicate that 5 of the disputed texts were written by Hamilton and the other 7 by Madison. However, to this point in our analysis we have no reliable way of calculating the probability of group membership for each paper.

3.3. THE BOOTSTRAP

The bootstrap (Efron, 1979; Efron *et al.*, 2001) solves this problem by enabling a numerical estimate of the standard error of test statistics for a group. Essentially, the bootstrap provides for comparisons of individual scores to groups of sample scores, allowing for accurate probability estimates of group membership.

In our case, we used the bootstrap to compare the logistic function scores of the individual *Federalist* texts to the function scores of the three groups of texts – Hamilton, Madison, and disputed. The bootstrap enables these comparisons by generating a large number of sample sets, each consisting of randomly drawn and reconstructed replacements from the original data set (Efron *et al.*, 2001). In other words, the bootstrap uses a randomized selection of scores from the papers in a group to generate a new group of the same size. The bootstrap repeats this procedure many times (in our case 1,000 times) to create a large number of sample groups. It is from this large set that the test statistics are then available for making accurate comparisons.

One of the primary benefits of using the bootstrap is that probability estimates may be made using *any* of the available test statistics (e.g., standard deviation, median, mode) rather than being limited to using the mean (Efron *et al.*, 2001). In our analysis, we made the comparisons using the median as indicator of each group's central tendency. We a priori chose the median as the best indicator because it is relatively insensitive to extreme scores in a distribution.

We used the bootstrap to generate three distributions ($B = 1,000$ samples) of median values, one distribution for each of the three groups. These large distribu-

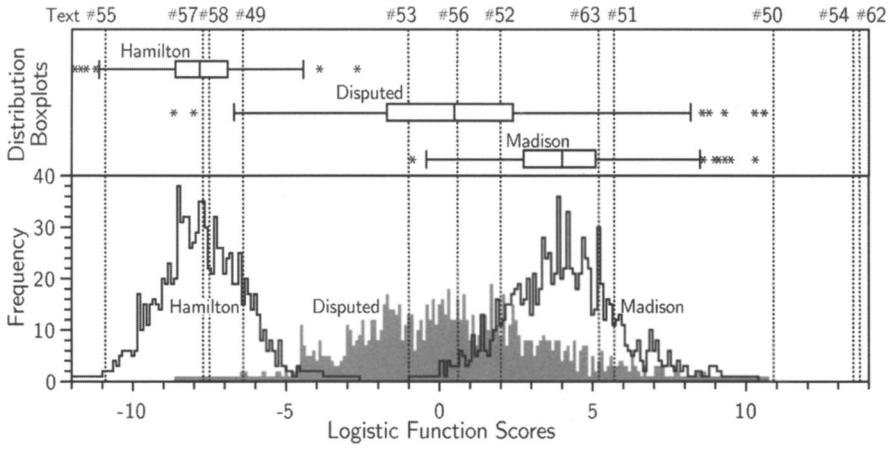

Figure 1. Frequency plot and boxplots of the logistic function scores of the bootstrapped groups ($B = 1,000$ samples per group). The dashed vertical lines represent the individual logistic function scores of the disputed papers.

Table VI. Logistic function score of each of the 4 undisputed papers misclassified by discriminant analysis. The scores are compared to the bootstrap distributions (B = 1,000 per group)

Paper (Author)	Logit score	Z-score (probability of group membership)		
		Hamilton	Madison	Disputed
14 (Madison)	–1.36	4.96 (< 0)	–3.12 (< 0)	–0.63 (52.86)
41 (Madison)	–1.09	5.17 (< 0)	–2.96 (0.30)	–0.54 (58.92)
46 (Madison)	–1.24	5.05 (< 0)	–3.05 (0.22)	–0.59 (55.52)
23 (Hamilton)	1.49	7.15 (≪ 0)	–1.47 (14.16)	0.34 (73.38)

tions provide for confidence intervals around the test statistic (median) for each of the groups. Figure 1 is a graphic representation of the result of the bootstrap and Table V provides the summary information for these groups. With this procedure complete, we could compare the logistic score of the individual papers to the bootstrap distributions. This allows a calculation of the probability of an individual paper score falling within each bootstrap distribution of scores.

To find out whether or not the bootstrap improved the efficacy of the discriminant function, we returned to the papers that were misclassified during the posterior probabilities test (Table IV). Comparing the logistic function score on each of these misclassified texts to the bootstrap distributions yields the results shown in Table VI.

The scores of the three misclassified Madison papers are outliers from both authors, but each paper falls more confidently within the Madison distribution than in the Hamilton distribution. On the other hand, the single misclassified Hamilton

Table VII. Logistic function score of each disputed paper compared to the bootstrap distributions ($B = 1,000$ per group)

Paper	Logit score	Z-score (probability of group membership)		
		Hamilton	Madison	Disputed
49*	–6.44	1.04 (29.84)	–6.06 ($\ll 0$)	–2.35 (1.88)
50	10.86	14.39 ($\ll 0$)	3.96 (< 0)	3.52 (< 0)
51	5.70	10.41 ($\ll 0$)	0.97 (33.20)	1.77 (0.08)
52	2.04	7.58 ($\ll 0$)	–1.15 (25.02)	0.52 (60.30)
53	–1.03	5.21 (< 0)	–2.93 (0.34)	–0.52 (60.30)
54	13.47	16.40 ($\ll 0$)	5.47 (< 0)	4.40 (< 0)
55*	–10.93	–2.43 (1.50)	–8.66 ($\ll 0$)	–3.88 (< 0)
56	0.62	6.48 ($\ll 0$)	–1.97 (4.88)	0.04 (96.80)
57*	–7.66	0.10 (92.04)	–6.77 ($\ll 0$)	–2.77 (0.56)
58*	–7.51	0.21 (83.36)	–6.68 ($\ll 0$)	–2.72 (0.66)
62	13.71	16.58 ($\ll 0$)	5.61 (< 0)	4.49 (< 0)
63	5.24	10.05 ($\ll 0$)	0.70 (48.40)	1.61 (10.74)

* = texts closer to Hamilton than Madison group.

paper remains misclassified because its score falls much more confidently into the Madison group.

So, using the bootstrap in conjunction with the discriminant function results in the correct attribution of 98.0% of the Hamilton texts (50 of 51) and 100.0% of the Madison texts (14 of 14). Recall that scores produced by the logistic function represent the combination of representational language features tagged within each text. The results of this analysis engender confidence in the discriminatory power of the logistic function and they also demonstrate the potential usefulness and sensitivity of our rhetorical theory.

You will notice the four misclassified papers fall most confidently into the bootstrap distribution of disputed texts (the rightmost column of Table VI). We believe this finding, along with the misclassification of these papers by the discriminant function, are actually indicative of the particular collaborative process under which these texts were written, an argument we will more fully develop in the next section.

Finally, we turned to the disputed papers and, using the same comparison procedure, computed the probability each disputed paper's score would fall into the bootstrap distributions. Table VII is the result. Our analysis indicates four of the disputed papers are most confidently attributed to Hamilton and the remaining eight to Madison.

These findings differ markedly from the results of Mosteller and Wallace's (1964) function word investigation, but they are consistent with the results of other

investigations (Martindale and McKenzie, 1995; Holmes and Forsyth, 1995). We believe we know why: a bit of rhetorical theory will help with our explanation.

4. Discussion

The Roman scholar Cicero compiled and extended the rhetorical thinking of his day into his treatise, *Ad Herennium* (1954). Cicero was interested in what today we might label a cognitive approach to teaching rhetoric, carefully delineating the five activities of a speaker: invention, arrangement, style, memory, and delivery (I.ii.2–3, cf. May and Wisse, 2001, pp. 29–32). Our contemporary rhetorical theories of writing focus most heavily on invention, a process that, as Cicero suggests, is employed throughout the composition process to attain a desired end (I.iii.4). On the other hand, many of the techniques of author attribution, including the function word analysis employed in the Mosteller and Wallace (1964) study, concentrate on what rhetoricians would recognize as style or delivery – the way the words and sentences come together and fall onto the page. Most attribution methods purposely focus on these aspects, finding and quantifying features of texts that are "relatively immune" from an author's conscious control (Bailey, 1978; Holmes, 1994, pp. 87–88).

Our attribution method differs because the representational theory behind our tagging software takes a broader view of Cicero's activities, attempting to account for how the activity of invention (where rhetorical plans are hatched, continuously revised, and ultimately linked to linguistic patterns) manifests itself in the other activities leading to the final text. In other words, our characterizations suggest how the author has reacted linguistically, both consciously and unconsciously, to the particular rhetorical situation.

Both Hamilton and Madison were gifted speakers and writers. Both demonstrated great expertise with the language and could vary their methods of persuasion. Nonetheless, there were some noted differences in ability: Hamilton, known as "a literary craftsman of the highest order" (Adair, 1944, p. 241), could rely on his excellent command of language to excite audience reaction. Madison, also an accomplished political pamphleteer, was known for his deep historical understandings and copious notes of all the arguments made during the Constitutional convention (Adair, 1944, pp. 111 and 249–250) and could, therefore, bring the specific lessons of the past alive for his contemporary audience. These differences in ability would affect the decisions the author would make as he wrote. So would the particular subject of an argument. For example, it would be difficult for an author to use examples for or against a new plan of taxation that simply has no precedent from which to argue. Instead, the author would have to speculate on how the system would work or would have to try to compare the new plan to a somewhat similar plan. This goes to show some ways the activity of rhetorical invention is affected by the rhetorical situation in which a text is composed.

The rhetorical situation in which Madison and Hamilton found themselves had three unique features that are important for our discussion. First, neither author at the time he was composing wanted to be identified: they shared the goal of getting the Constitution ratified and, therefore, both saw the need to enable "Publius" to take positions of compromise that were politically untenable for the authors individually (Adair, 1944; Carey, 1984; Kesler, 1999). Second, both authors faced deadlines and time pressures from their other responsibilities (Adair, 1944, pp. 239–241) which led them toward satisficing behavior in their composition processes, accepting results that may not have been ideal, but that would suffice. In the preface to the first collected edition of *The Federalist*, Hamilton felt the need to apologize for the "violations of method and repetition of ideas which cannot but displease a critical reader" (qtd in Adair, p. 241). Third, the authors shared an explicitly collaborative situation. Although both maintained the overall goal for the project, each had his intellectual and political axes to grind and wanted the shape of the future government to reflect his own thinking (e.g., on questions of taxation Adair, 1944, pp. 250–251).

Addressing ourselves to the nature of the *collaborative* effort, what emerges from the Table VII results is a more ambiguous picture of authorship than the one which warrants many attribution studies. That is, accepting the ambiguity of the authorship question (rather than declaring one or the other author as *the* culprit) reveals a picture that not only accounts for a sophisticated understanding of the collaborative nature of rhetorical invention, but also one more congruous with history.

Take, for example, paper 53, a text that cannot be very confidently placed with either author's group. Douglass Adair cites evidence of sentences and whole paragraphs that could be matched to Madison's earlier writings (p. 117), but then also points to paragraphs in the paper discussing topics that Hamilton had clearly "made his own" in earlier papers (p. 119). And consider papers 57 and 58, both of which score solidly as Hamilton texts and both of which were written just at the time Madison was being desperately urged to return to Virginia (pp. 253–254). The historical record does not preclude the possibility that a hurried Madison turned to his collaborator for help getting these finished and to the printer for him.

Unfortunately, the specifics of the collaborative effort of the individuals writing each paper are lost to time. What is clear to us, however, is that the answer to the authorship question of the disputed papers is more complicated than has been commonly presumed since the publication of Mosteller and Wallace's overwhelming function word results. We would suggest that our evidence, coupled with the historical accounts of the project, the nature of rhetorical invention (see especially LeFevre, 1987), and the findings in other authorship studies supports a different understanding of what it means to be an author in a collaborative situation than the atomistic authorship often presumed to be the case.

When we consider the results shown in the right-hand column of Table VII, our findings suggest a discernible difference in the composing process for these

disputed papers. While this certainly does not prove or disprove the proposition that Madison penned the disputed texts, it does suggest that findings like ours and those of other investigators who have not achieved the definitive results of Mosteller and Wallace's famous analysis (e.g. Martindale and McKenzie, 1995), are not necessarily aberrant. Instead, they may be pointing to an essential element in the composition process: the effects of collaboration on texts produced in collaborative rhetorical situations.

Revisiting each of the misclassified papers from the posterior probability analysis (Table IV) illustrates our point more clearly. Hamilton's paper 23 was written to begin the third formal section of *The Federalist* project. Adair (p. 250) recounts that after the publication of paper 22, the authors agreed to apply a different division of labor, dividing the papers by topic, rather than by individual papers as they had been doing. It likely would have been during times of transition in the project that the two men would, as Kesler puts it, "consult with one another on the general direction of the series and the division of labor emerging within it" (1999, p. xiii). Such consultations would tend to reenergize a collaboration, reminding the participants of their need to adjust their own rhetorical activities, especially invention, to accommodate the larger aim of the project. Such transitions between sections would be where an author would need to concern himself with the overall flow of the project and with the sensibilities of the other participants.

We see in this misclassified Hamilton paper the possibility that authoring under more explicitly collaborative conditions makes for a different rhetorical situation and, thus, a different result. Similarly, Madison's papers 41 and 46 begin and end a series of six papers he authored, transitioning into and back out of an in-depth discussion of federalism and the powers of the national government. Likewise, Madison's paper 14 was written early in the project to conclude the first block of papers with a flourish. As can be seen in the right-hand column of Table VI, all four misclassified papers fit most confidently into the "disputed" group of papers.

5. Conclusions

Our findings lead us to conclude that different intensities and types of collaboration between the authors took place across *The Federalist* papers. Since collaboration is surely a part of the rhetorical situation, collaboration impacts an author's rhetorical activities – especially invention – in ways that are discernible from the representational language strings in the resulting texts. The effects of collaboration in *The Federalist* project may account for why both authors felt and claimed authorship for the disputed papers and may also account for the difficulties investigators have had achieving the same results as Mosteller and Wallace, who chose function words specifically based upon the words' disparate usages in the papers.

From our results, we cannot conclude that Madison did or did not pen the disputed papers. However, we feel confident that since *The Federalist* was a collaboratively written project, the common notion of atomistic authorship is suspect

and, as the results of our analyses indicate, problematic. More research is needed to better understand the effects of collaboration on writing processes. A key difficulty with an historical exploration like ours is that the authors are not available to interview about their rhetorical activities. Studies analyzing the works of authors in contemporary natural or experimental settings to see how works differ when written with varied implicit and explicit forms of collaboration could provide confirmatory evidence of our findings.

We feel our study helps confirm the viability of a representational theory of composition that is based upon writers' choices of language in service of rhetorical effect. This suggests potentially important implications for fields related to computational linguistics. Any computer system that purports to communicate richly with human readers or to "comprehend" the natural languages produced by humans must, somehow, account for the relationships between rhetorical plans and resulting textual patterns. We feel representational composition theory provides a step in the direction of understanding these relationships. However, another important direction is to see whether and how this rhetorical theory applies beyond the narrow linguistic bounds of standard written English. For example, it is uncertain what dimensions of language effect would be required to attempt to characterize the rhetorical plans produced by communicators in other linguistic traditions.

We see many possibilities for multivariate exploration of texts using our language theory as the basis for analysis. Already we are using factor analyses of text corpora to develop a different, detailed understanding of textual genre and how it relates to representational language usage (cf. Biber, 1988). We are also exploring relationships between teacher comments on student writing and the students' subsequent revisions to see how teachers intervene in their students' activities of rhetorical invention. We anticipate many other studies that could be carried out using our automated tagging and visualization software.

As regards *The Federalist*, our findings do not definitively solve the authorship problem. Instead, our findings suggest that the nature of the problem as it has been posed is itself a problem because it is premised upon a notion of atomistic authorship that does not hold in all situations, especially in situations that are collaborative.

Acknowledgements

The views expressed in this paper are those of the authors and do not reflect the official policy or position of the U.S. government or the Department of Defense. The authors thank Barbara Johnstone, Beverly Sauer, Joseph Rudman, and two anonymous reviewers for their insightful comments on an earlier draft of this paper.

Appendix

A. Description of the 18 Representational Effects

Our string catalog is used as part of our writing classroom and, therefore, evolves. At the time of this writing it characterizes over 250 million strings according to 18 non-overlapping categories of representational effect. Below is a brief description of the effects, grouped into three clusters: *thought*, *description*, and *linear cues*. A fuller description of this language theory may be found elsewhere (Kaufer and Butler, 2000).

Each description is followed by phrases containing a few underlined strings that are assigned to the representational effect.

A.1. Effects of the *thought* cluster:

First Person: Often when reading we get an impression of a mind at work, either the author's or a narrator or other character. Writers often use strings containing self-referential pronouns to individuate the point of view of a single consciousness from objects outside of it. Example strings: I have peaches in my cart and I'm positive they're organic.

Inner Thinking: Private and particular minds on the page don't necessarily require first person. Writers can use English strings that give readers the impression of private minds at work by using, for example, thinking verbs or assurances. Example strings: Never did we imagine the need to heed the meaning of "organic."

Think Positive: Writers signal positive feeling in the worlds they create by using strings that contain "feel good" words and phrases. These strings are less constrained than *inner thinking* strings because they can be unassociated with more complete thoughts. Example strings: Recent laws have made it easier to know what you put in your cart – a welcome change that promises many a new relationship with organic foods.

Think Negative: Likewise, writers have available a large class of strings evoking distress in the mind of the writer, narrator, or character. Even wrapped around neutral descriptions, strings of *negative affect* signal the reader that the writer disapproves. Example strings: It was those abuses that produced the restraints that prohibit the inspection practices overseas.

Thinking Ahead: This dimension is evident when a text contains projections into the future. These strings capture a mind anticipating a potential event rushing toward the present with a prospect of actualization. Example strings: The government will get into oversight because no bureaucrat will want to be blamed for missing a problem.

Thinking Back: Thinking-back language effects occur when a text contains retrospections on a past. The reader feels a mind recollecting an event that had assumed or experienced actuality and that is now receding into memory. Example strings: The legislation has made it easier and may have prevented a problem. The old law was to have kept shoppers forever guessing about the "organic" label.

A.2. EFFECTS OF THE *description* CLUSTER:

Word Picture: Writers use these strings to prime images that embody all the major elements of the story. Writers prime a word picture to allow readers to "see" the skeleton of the story in mental imagery. Example strings: It set about hiring 100 analysts in 56 cities across Europe.

Space Interval: These strings prime the reader's sense of spatial contiguity. English relies primarily on a handful of strings containing prepositions (e.g. on, under, across, against, over, alongside, in front of, in back of) to carry the burden for signaling relationships between objects occupying contiguous space. Example strings: It will share violations with news agencies, including the *Times*, he added, saying a new press office will be built near the Brandenburg Gate.

Motion: These strings prime kinetic images in the mind of the reader. They force readers not only to build an image of motion, but also to specialize the kinetic shape as part of their basic understanding of the world created by the writer. Example strings: France's Renseignements Generaux can open mail and tap farmers' phones at will.

Past Events: The simplest way for readers to feel the elapse of time is through strings containing the simple past tense. The advantage of the simple past for writers is that event starts and stops can be signaled to the reader. Example strings: They just caught a case of fraud as the agency got set up and operated out of Germany.

Shifting Events: Another way English conveys time's passage is by creating shifts across events. These shifts, often captured in strings containing adverbials, occur in both time and space. In the physical world, time and spatial shifts invariably co-occur. However, English phraseology often separates shifts in time from shifts in space, providing writers with rich access to time adverbials and space adverbials that do not overlap. Example strings: Once food enters the country it will be labeled at the same time it is inspected.

Time Interval: Event uniqueness or repetition is usually indicated by writers through strategic selection of strings containing time adverbials. Temporal

adverbs are often used in repeated event strings. Beyond single-word adverbs, writers of English encode repeated events through a large inventory of adverbial and prepositional strings, all signaling temporal recurrence. Example strings: The agency is already changing. The last time it got involved was during the cold war years.

A.3. EFFECTS OF THE *linear cues* CLUSTER:

Cue Common Knowledge: Writers use these language strings to cue a reader's prior knowledge of standards, resemblances, authorities, precedents, and values. The ancient rhetoricians had an umbrella name for this class of priors – commonplaces. Commonplaces increase the solidarity between writer and reader, as merely referencing a commonplace highlights shared knowledge, beliefs and values that the writer can use as leverage for persuasion. Strings cueing prior knowledge make communication with readers more efficient, as commonplaces function as implicit premises to carry conclusions. Example strings: The food will be safe only when security and liberty are assured.

Cue Prior Text: Writers increase a reader's sense of familiarity by cueing information that readers have learned in the course of reading the text. Such cueing provides important grounding to readers within the text. Readers understand that the writer has composed the text to take into account the familiar information that results from the reader's progress through the text. Example strings: Does this sound familiar? To avoid more of them the agency needs oversight to let it know its boundaries.

Cue Reader: Using these strings, the writer acknowledges – in some cases fully addresses – the reader's presence, guiding the reader explicitly or previewing coming features of a text. These are the chief strings through which writers indicate their intentions and purposes to readers as well as telegraph the plan of the text to come. Example strings: Does this seem silly? Remember, we are shielded by our laws and protections.

Cue Notifier: Writers use these language strings to signal readers about key terms and information, indicating the presence of chunks of information to provide readers with a key discriminator in helping them create schemata and conveying the text's organization. Example strings: The paradox is that the agency would be a different organization, a kind of food intelligence agency.

Cue Movement: These language strings address a reader who is negotiating a physical action in his or her proximate environment. The action may be related to some institutional procedure or practice (like filling out a tax form), may require mental focus and review, or may direct physical actions with the reader's hands or body. Example strings: To ensure organic purity put the food

Table VIII. The variables (categories of representational effect) to which each of the tagged strings in the first sentence of disputed text 49 was assigned

Variable	Matched strings
Inner Thinking	[expected to]
Think Positive	[that valuable]
Thinking Ahead	[in order to]
Thinking Back	[had been prepared]
Past Events	[quoted in] [be called in]
Cue Notifier	[, which] [, by]
Cue Reasoning	[, for the]

labels under more scrutiny. Rotate the package quickly and look for the new EU holographic symbol.

Cue Reasoning: Using these strings, writers guide the reader as a thinking being who, in following the text, follows a path of reasoning. Strings of this type launch lines of reasoning the reader is meant to follow. This class is marked by language strings indicating logical connection. Example strings: Nor, for that matter, has the government, but even if industry needed nothing else, inspections might falter.

B. Example Tagged Sentence

The text below is the first sentence from *The Federalist* paper number 49, one of the papers with disputed authorship. The underlined strings are those matched by the tagging software:

> THE author of the "Notes on the State of Virginia," quoted in the last paper, has subjoined to that valuable work the draught of a constitution, which had been prepared in order to be laid before a convention, expected to be called in 1783, by the legislature, for the establishment of a constitution for that commonwealth.

The nine strings tagged in the example are categorized as shown in Table VIII. Notice some of the representational strings include punctuation marks in addition to words.

Figure 2 shows the text tagged by our software, DocuScope. A larger color version of the figure is available on the internet (Collins, 2001). When using the software, the color of each underline indicates to which of the 18 mutually exclusive representational effects the tagged string has been assigned. The variable

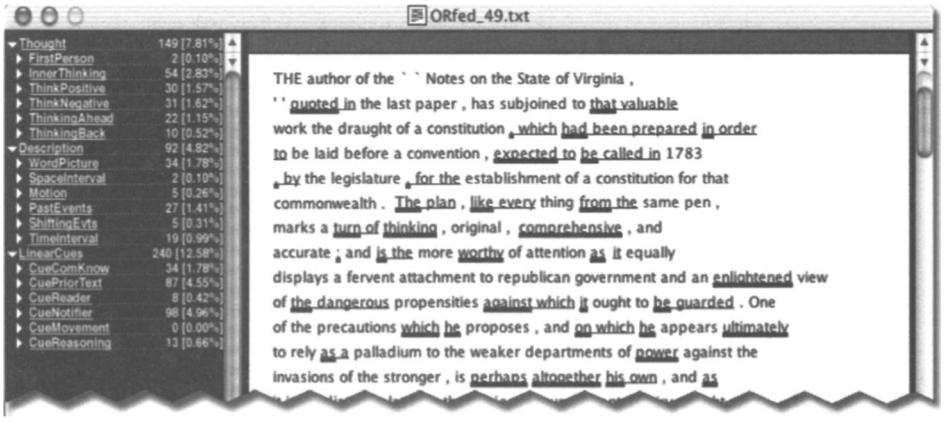

Figure 2. Detail showing the tagged text on screen.

Table IX. The logistic discriminant function and example text score

Variable	Coefficient	Example score (Text #49)
First Person	–15.2857	0.10
Think Positive	5.3225	1.57
Think Ahead	–4.9676	1.15
Think Back	0.4421	0.52
Motion	–20.7707	0.26
Past Events	19.1744	1.41
Shifting Events	13.6757	0.31
Time Interval	–7.3426	0.99
Cue Prior Text	–0.4287	4.55
Cue Reader	–4.1423	0.42
Cue Notifier	1.5517	4.96
Intercept	–30.3952	

scores for the tagged *Federalist* paper (seen at the left side of the figure) may be saved to a comma-delimited text file for analysis.

C. Logistic Discriminant Function

In *The Federalist* case, our criterion variable is *author*: each text will be assigned either to the group of texts authored by Hamilton or to the group by Madison. The assignment will be based upon the scores of the individual text's response variables. We describe the development and validity of this function (including the

use of a stepwise procedure) in section 3.2 of the paper. The formula yielded by this approach is described by Table IX.

The logistic score for a text is calculated by multiplying a text's score on each response variable by that variable's coefficient, summing these products and then adding the intercept value. For example, the logistic function score of *Federalist* paper 49 (the first disputed paper) is calculated by summing the products of the text's variable scores (the right-hand column of Table IX) and the variables' coefficients with the intercept value (−30.3952). For paper 49, this yields a logistic function score (or "logit") = −6.438. Since this score is less than zero, this disputed paper is assigned by the discriminant function to Hamilton. Papers scoring higher than zero would be assigned to Madison. The logit score for each disputed text is found in column 2 of Table VII.

References

Adair D. (1944) The Authorship of the Disputed Federalist Papers (2 Parts). *William and Mary Quarterly*, 3rd Series, pp. 97–122 and 235–264.

Bailey R.W. (1978) Authorship Attribution in a Forensic Setting. In Ager D.E., Knowles F.E., and Smith J. (eds.), *Fifth Internat'l Symposium on Computers in Literary and Linguistic Research*. U of Aston, Birmingham, pp. 1–20.

Biber D. (1988) *Variation Across Speech and Writing*. Cambridge University Press, Cambridge.

Carey G.W. (1984) Publius – A Split Personality? *Review of Politics*, 46(1), pp. 5–22.

Cicero (1954) *Ad Herennium: On the Theory of Public Speaking*, Loeb Classical Library. Harvard UP, Cambridge, MA.

Collins J. (2001) Collaborations in Text: *The Federalist* Case Support Site. [http://betterwriting.net/projects/fed01/ [10/30/02]].

Ede L., Lunsford A. (1990) *Singular Texts/Plural Authors: Perspectives on Collaborative Writing*. Southern Illinois University Press, Carbondale.

Efron B. (1979) The Bootstrap. In *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics, Philadelphia, pp. 27–36.

Efron B., Rogosa D., Tibshirani R. (2001) Resampling Methods of Estimation. In Smelser N. and Baltes P. (eds.), *International Encyclopedia of Behavioral and Social Sciences*, Vol. 19. Elsevier Science, pp. 13216–13220.

Fellbaum C. (1998) *WordNet: An Electronic Lexical Database, Language, Speech, and Communication*. MIT Press, Cambridge, MA.

Fienberg S.E. (1979) *The Analysis of Cross-classified Categorical Data*. MIT Press, Cambridge, MA.

Holmes D.I. (1994) Authorship Attribution. *Computers and the Humanities*, 28(1), pp. 87–106.

Holmes D.I., Forsyth R.S. (1995) The Federalist Revisited: New Directions in Authorship Attribution. *Literary and Linguistic Computing*, 10(2), pp. 111–127.

Hosmer D.W., Lemeshow, S. (2000) *Applied Logistic Regression*. John Wiley and Sons, New York.

Irizarry E. (1993) The Two Authors of Columbus' Diary. *Computers and the Humanities*, 27, pp. 85–92.

Kachigan S.K. (1991) *Multivariate Statistical Analysis*, 2nd edition. Radius Press, FDR Station, NY.

Kaufer D.S., Butler, B.S. (1996) *Rhetoric and the Arts of Design*. Lawrence Erlbaum Associates, Mahwah, NJ.

Kaufer D.S., Butler, B.S. (2000) *Principles of Writing as Representational Composition: Designing Interactive Worlds with Words*. Lawrence Erlbaum Associates, Mahwah, NJ.

Kaufer D.S., Carley K.M. (1993) *Communication at a Distance: The Influence of Print on Sociocultural Organization and Change, Communication*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Kesler C.R. (1999) Introduction to The Federalist Papers. In Rossiter C. (ed.), *The Federalist Papers*. Mentor, New York, pp. vii–xxxiv.

LeFevre K.B. (1987) *Invention as a Social Act*. Southern Illinois University Press, Carbondale.

Mallon T. (2001) *Stolen Words: The Classic Book on Plagiarism (Updated)*. Harcourt, New York.

Marcu D. (2000) The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach. *Computational Linguistics*, 26(3), pp. 395–448.

Martindale C., McKenzie D. (1995) On the Utility of Content Analysis in Author Attribution: The Federalist. *Computers and the Humanities*, 29(4), pp. 259–270.

May J.M., Wisse J. (2001) Introduction. In *Cicero: On the Ideal Orator (De Oratore)*. Oxford University Press, New York, pp. 3–56.

Mosteller F., Wallace D.L. (1964) *Inference and Disputed Authorship: The Federalist*. Addison-Wesley Publishing Company, Reading, MA.

O'Hara T.F., Hosmer D.W., Lemeshow S., Hartz S.C. (1982) A Comparison of Discriminant Function and Maximum Likelihood Estimates of Logistic Coefficients for Categorical-scaled Data. *Journal of Statistical Computation and Simulation*, 14, pp. 169–178.

Richards I.A. (1991) Context Theory of Meaning and Types of Context. In Berthoff A.E. (ed.), *Richards on Rhetoric*. Oxford University Press, New York, pp. 111–117.

Rudman J. (1998) The State of Authorship Attribution Studies: Some Problems and Solutions. *Computers and the Humanities*, 31, pp. 351–365.

Rudman J. (2000) Non-Traditional Authorship Attribution Studies: *Ignis Fatuus* or Rosetta Stone?. *Bibliographical Society of Australia and New Zealand Bulletin*, 24(3), pp. 163–176.

Scolari (2000) Diction 5.0: The Text Analysis Program. [http://www.scolari.co.uk/diction/diction.htm [6/26/01]].

Seber G.A.F. (1984) *Multivariate Observations*. John Wiley and Sons, New York.

Stillinger J. (1991) Multiple Authorship and the Question of Authority. *Text: Transactions of the Society for Textual Scholarship*, 5, pp. 283–293.

Stone P.J. (2000) Comparison of General Inquirer with Other Text-analysis Procedures. [http://www.wjh.harvard.edu/~inquirer/ [4/4/01]].

Tukey J.W. (1949) Comparing Individual Means in the Analysis of Variance. *Biometrics*, 5, pp. 99–114.

Turner D. (1992) Project Gutenberg Etext of The Federalist Papers. [http://lcweb2.loc.gov/const/fed/fedpaper.txt [1/18/01]].

Vlachos P. (2001) StatLib Datasets Archive. [http://lib.stat.cmu.edu/datasets [10/21/02]].

Wilks S.S. (1932) Certain Generalizations in the Analysis of Variance. *Biometrika*, 24, pp. 471–494.

Woodmansee M. (1994) Genius and the Copyright. In *The Author, Art, and the Market: Rereading the History of Aesthetics*. Columbia University Press, New York, pp. 35–55.