The diary of a public man: a case study in traditional and non-traditional authorship attribution

David I. Holmes and Daniel W. Crofts
The College of New Jersey, Ewing, USA

Abstract

Of all American literary mysteries *The Diary of a Public Man* has been perhaps the most perplexing. Set principally in Washington DC, it covers a short but critical period in the nation's history, the secession winter of 1860–61. The Diary entries are dated during the last months of James Buchanan's ill-fated administration and the first 2 weeks after Abraham Lincoln's inauguration. The publisher refused to name the author yet, despite the Diary's anonymity, it has been used and quoted by historians for more than a century. It is clearly the work of an exceptionally gifted writer. The Diarist pictures himself as a strong Union man, much worried whether the crisis can be resolved without resort to war. Naturally there has been much speculation as to the Diary's authorship. This article describes how traditional and non-traditional methods of authorship attribution may be employed on the Diary, which we believe to have been written by William Henry Hurlbert. We argue that the joint interdisciplinary approach employed in this article should be the way in which attributional research is conducted. Information on the traditional attribution section of this article is adapted from the forthcoming book by Daniel W. Crofts, *A Secession Crisis Enigma: William Henry Hurlbert and 'The Diary of a Public Man'* (Louisiana University Press, 2010)

Correspondence:
David I Holmes, Department of Mathematics and Statistics, The College of New Jersey, 2000 Pennington Road, Ewing, NJ 08628-0718, USA
E-mail:
dholmes@tcnj.edu

1 Introduction

In 1879 the *North American Review* published four separate monthly installments of *The Diary of a Public Man*. The purported Diary attracted wide interest. Although the editor concealed the diarist's identity and deleted the names of some of the persons he mentioned, the document amply fulfilled the editor's promise to shed light on the 'dark and troubling times' during the months before the start of the American Civil War (Anonymous, 1879, introduction; Crofts, 2010, appendix). The Diary appeared to offer verbatim accounts of behind-the-scenes discussions at the very highest levels during the greatest crisis the country had ever faced.

The diarist's glimpses of several key principals, most notably Abraham Lincoln, have often reappeared in the writings of historians. On three occasions, the diarist reported having met directly with Lincoln, and he gathered additional second-hand reports of other conversations with him. All of these episodes took place during the weeks and days immediately before and after inauguration day,

Literary and Linguistic Computing, Vol. 25, No. 2, 2010.
doi:10.1093/llc/fqq005 Advance Access published on 14 April 2010

4 March 1861, a time of extraordinary importance. In late February and early March peace hung by a thread, as the new President confronted the terrible reality that seven states in the Deep South had begun to form a separate government. The diarist provided memorable details of his encounters with Lincoln. He detected 'something almost woman-like in the look of his [Lincoln's] eyes' and thought him 'the most ill-favored son of Adam I ever saw'. On the other hand, the diarist noted the 'quaint and rather forcible way' that Lincoln expressed himself, often using a homely analogue or telling a little story to make a point (Anonymous, 1879, entries for 20 February, 4, 7 March 1861).

The diarist had access to many others besides Lincoln. He reported meeting with a wide spectrum of key officials, from the South as well as from the North. The first Diary entry, dated 28 December 1860, recounted a 'long conversation' with James L. Orr, once a Speaker of the US House of Representatives and then a Commissioner from South Carolina, who impressed the diarist as 'honestly trying to make the best of what he felt to be a wretched business, and that at heart he was as good a Union man as anybody in Connecticut or New York'. The diarist was also close to Senator Judah P. Benjamin of Louisiana, whom he considered blessed with 'a rare and lucid intelligence' and 'alertness and accuracy of mind', but deficient in 'consistency of purpose and strength of will'. The same diarist who was in direct touch with Southern luminaries and disinclined to assign them full blame for the crisis also knew their polar opposite, Charles Sumner, the Republican senator from Massachusetts, who was a living symbol of Northern contempt for the slave system and its defenders. Although the diarist 'never affected an admiration' for Sumner, the latter allegedly sought him out in hopes of using his presumed influence with Lincoln (Anonymous, 1879, entries for 28 December 1860, 13 January, 3 March 1861).

Even the most casual reader will find the Diary accessible and the narrative lively. The diarist had an uncommon knack for capturing quickly the essence of a situation. His account is punctuated with embedded quotations that convey immediacy. Occasionally, the Diary's daily entries record only a single significant conversation or episode. More often than not, however, the diarist summarizes conversations with several persons, frequently punctuated by his own reactions, opinions, or observations. He reports encountering people on the street, staying out late, receiving mail, and being 'called out by a card' from a visitor. The entries unfold as the diarist recalls the events of the day. In some entries the diarist almost talks to himself—he heard a story 'so characteristic of all the persons so concerned in it that I must jot it down'. The Diary therefore reads like a diary (Anonymous, 1879, entries for 25, 26 February, 3, 4, 9 March 1861).

From the Diary's first publication in 1879 until the late 1940s, most historians considered it authentic. They kept trying to identify the author, thereby continuing a chase that began the moment it first appeared, and they were baffled by their inability to pinpoint anyone. A few harboured doubts about the Diary, but leading scholars of the late nineteenth and early 20th centuries such as James Ford Rhodes, Frederic Bancroft, and Allen Johnson each made use of it. Bancroft and Rhodes corresponded with each other about the diarist's possible identity. 'I hope that we can yet solve this mystery', Bancroft wrote to Rhodes in 1896, 'for, I do not believe the diary was bogus' (Bancroft, 1896). Many of their successors of the 1930s and 1940s—most notably, David M. Potter, Allan Nevins, and Roy Franklin Nichols—also incorporated information from the Diary in their books. Potter quipped in 1942 that historians always quoted the Public Man's 'lively passages', even when they questioned his existence (Potter, 1942, p. 385).

In 1948, however, Frank Maloy Anderson (1871–1961) published *The Mystery of 'A Public Man': A Historical Detective Story* (Anderson, 1948). His book posed the two key questions regarding this remarkable document—who was this diarist? Was the diary genuine? Anderson's conclusions were buttressed by two decades of historical detective work. After Anderson's death, his research materials were deposited in the Library of Congress, thereby making it possible to examine the evidence and his assessment of it.

Anderson concluded that the author was Samuel Ward (1814–84), a talented and cosmopolitan

New Yorker, the brother of Julia Ward Howe ('Battle Hymn of the Republic'). Ward was best remembered for his emergence after the war as a uniquely successful Washington insider, 'The King of the Lobby'. Anderson explained that he fixed upon Ward by specifying a number of key characteristics the diarist must possess, and by finding that Ward matched the criteria far better than any other possible diarist. The author had to be someone who was resident in Washington during the secession winter, who enjoyed 'extensive personal acquaintance' with political leaders there and had 'considerable knowledge [of] and interest in politics', but who was not himself someone with strong partisan affiliations. He must also be someone of 'wide experience' who was well informed about business affairs. He must either be 'a New Yorker' or someone thoroughly acquainted with prominent New Yorkers. On the key issue of the day, the diarist must have 'distrusted and feared the influence of extremists on both sides of the crisis' and had 'a manifest preference for moderate men'. In particular, he needed to be someone 'who was upon at least fairly intimate terms with [Secretary of State William H.] Seward, for whom he had a qualified admiration, and with [Illinois Senator Stephen A.] Douglas, for whom his admiration was in every respect deep and sincere' (Anderson, 1948, p. 140–61, quotations on 140–44).

Although satisfied that he had finally identified the elusive diarist, Anderson concluded that the Diary itself was a clever concoction rather than an actual document:

> It is not a genuine diary actually kept in 1860–1861. It is, on the contrary, in part genuine and in part fictitious. It includes as a core a genuine diary, probably rather meager, actually kept by Sam Ward at Washington during the Secession Winter of 1860–1861. Attached to this genuine core there is a large amount of embellishment added at a later date. This added increment is in part recollection and in part pure invention. The genuine core, the recollection, and the invention have all been skillfully blended with a polished literary style.

Consequently, Anderson concluded that the diary 'ought not to be regarded as a reliable source in any of its details' and it 'ought not to be regarded as history'. Its only redeeming value lay in the 'substantially accurate' impression it conveyed of the 'confusion and uncertainty' that enveloped Washington during the secession winter (Anderson, 1948, p. 169, 178).

2 'Traditional' Attribution

Frank Maloy Anderson judged that Sam Ward must have written *The Diary of a Public Man*. Anderson, however, had little to say about writing style. How does the Diary match up against prose written by Ward? Does it contain patterns of word usage that characterize Ward's writing? Textual analysis tells a different story than the one in Anderson's book.

The most extensive prose venture Ward ever undertook was a memoir of his experiences in the California gold fields in 1851–52. Ward's charming and evocative recollections were published in fourteen instalments in *Porter's Spirit of the Times*, a New York weekly, starting on 22 January 1861, and concluding abruptly with the issue of 23 April 1861, just after the war began. The series ended because Ward departed for the seceded states in the company of William H. Russell, the renowned correspondent for the London *Times* (Collins, 1949).

For purposes of analysis, therefore, we have an abundance of material that Ward wrote in early 1861. At precisely the same time the political crisis reached its most acute phase, he was crafting approximately 4,000 words per week on an entirely different subject. Do the Gold Rush memoir and *The Diary of a Public Man* match up? Do they read as if the same person wrote both? The answer to this question is beyond dispute. Use of traditional attributional tools of content and style shows that these two documents could not have been written by the same person.

It should be expected, of course, that the two documents would have dissimilar features. One is a purported Diary focused on elite level politics in 1861, even though it is now clear that the document was composed later. The other is a memoir of events that occurred almost a decade earlier, when Ward rusticated for over a year with a memorably diverse

group of miners, speculators, Mexicans, and indigenous peoples at an Indian encampment or 'rancheria' on the 'River of Grace', today's Merced River, which flows down from the high Sierras toward the San Joaquin River and San Francisco Bay.

The differences between the two documents go far beyond their radically different subject matter. Ward's memoir of the Gold Rush is awash with words having a '-tion' or '-ion' suffix, yet 6 of the 10 most favoured of these words do not appear at all in the Diary; three others are used but once; only 'determination' appears three times in the Diary. His sentences sometimes meandered in a Baroque manner. He had a habit of encasing unusual words or phrases within quotation marks and peppering his narrative with Spanish and French expressions. By contrast, *The Diary of a Public Man* was written by someone whose style featured active verbs accompanied by adverbs with an '-ly' suffix; indeed seven '-ly' words used in the Diary were absolutely central to the diarist's manner of written expression. Four of these ('really', 'perfectly', 'finally', and 'frankly') do not appear at all in the Gold Rush memoir, which is twice the length of the Diary; two are used once ('certainly' and 'entirely'); and one appears twice ('apparently'). Hardly any of the negatives used by the diarist appear in the memoir; Ward encountered no 'folly' and nothing that was 'wild', 'horrible', 'worse', or 'worst', and but single instances each of 'mischief' and 'wretchedness'. Ward did not precede his infrequent '-ly' adverbs with the word 'pretty', as was the diarist's habit. The diarist, unlike the memoir writer, found nothing 'mournful'. Both writers alliterated, and both used the words 'particularly', 'attention', 'determination', and 'anxiety', but each otherwise had his own distinct inventory of favourite words, and the differences far outweigh the similarities. The Diary is fast paced and immediate. The diarist, who supposedly recorded exchanges that took place that same day, more often provided the words from a conversation than did Ward's California memoir.

If the Diary and the memoir on the Gold Rush were not written by the same person, who besides Ward might fit Anderson's criteria for the diarist? This brings us to William Henry Hurlbert (1827–95), a gifted journalist and unconventional genius who covered his tracks so well that his role has been unrecognized until now. Hurlbert wrote with flair and found that he could make a good living with his pen. He became the chief editorial writer for the *New York Times* in 1857. By the time of the secession winter, however, Hurlbert and the *Times* had parted ways. Hurlbert moved to the *New York World* in 1862 and for 7 years starting in 1876 he became its editor-in-chief. At the *World* he brought his forceful, fluid writing style to bear on a wide range of topics—politics, international affairs, history, and literature.

Anderson knew that Ward and Hurlbert were close friends, and he recognized that 'the literary style of the Diary has a good deal of the pungency characteristic of almost everything that Hurlbert wrote'. He also surmised that the published Diary had been 'skilfully blended' by someone with 'a polished literary style'. The only role that Anderson could imagine for Hurlbert, however, was that of a possible collaborator, someone who might have helped enlarge Ward's 'genuine core' that dated to 1861. Hurlbert's 'lively imagination' could have 'been equal to the task of supplying inventions so plausible that they could pass for historical facts' (Anderson, 1948, p. 134–35, 169–70). Anderson was on the right track. He sensed that Hurlbert had applied his skills to the task at hand. He also knew that the finished document held together well. Rather than being episodic and fragmentary, as might be expected of a composite that had been cut and pasted together, it was strikingly coherent. Indeed, Anderson finally concluded that it was too good to be true. But he resisted the conclusion that we have reached in this attributional study, namely that Hurlbert wrote every word published in 1879. Whatever sources Hurlbert used to create it, the Diary was his work alone.

The Diary abounds with striking parallels to Hurlbert's distinctive writing style, some of which trace back to his very first years as a writer. His book about a trip to Cuba, published in 1854, made repeated use of key words that appear in the Diary—'wild', 'absurd', 'wretched', 'execrable', and 'sedulously', and it included frequent alliteration

(Hurlbert, 1854, p. 53, 56, 70, 83, 116, 119, 133, 167). His essay for the *Edinburgh Review* in 1856 included a number of the Diary's signature words—'certainly', 'singular', 'singularly', 'absurd', 'earnest', and 'earnestly' (Hurlbert, 1856). Hurlbert's editorials for the *New York Times*, written between October and December 1859, repeatedly anticipate patterns of usage that appear in the Diary. The editorialist leaned on words such as 'entirely', 'certainly', 'perfectly', 'apparently', 'especially', 'sedulously', 'earnestly', and 'clearly', all of which appear in the Diary and most of which stand out there. Both editorialist and diarist also employed 'deplorable', 'mischief', 'scheme', 'wild' and 'wilder', 'absurd', 'angry and agitated', 'astounding and alarming', 'dangerous and deplorable', 'dread and deprecate', 'dissolution and dishonor'—the list could be stretched to great length (Hurlbert, 1859).

As Hurlbert's career unfolded, he continued to write in ways that exhibit stylistic overlap with the Diary. When he escaped from Confederate captivity in August 1862, he described his experiences in a series of seven articles for the *New York Times*. These articles dovetail nicely with the Diary: both employ less than routine words such as 'forbear', 'ascertained', 'wildest', 'horrible', 'madness', 'wretched', 'pretensions', and 'peremptorily'; both include a great many words that begin with 'dis-'and 'in-'(Hurlbert, 1862). Writing editorials for the *New York World* in 1868, Hurlbert condemned the 'preposterous proceedings' of impeachment, castigated Republicans for heating 'the cauldron of public passions' and acting in 'hot and headlong haste', and predicted that their 'revolutionary recklessness' would lead to 'defeat and disgrace'(Hurlbert, 1868). In his 1874 editorials in defence of Henry Ward Beecher, Hurlbert complained that much of the evidence in the 'wretched business' was 'mischievous' and 'absurd'. It was 'perfectly plain' that the famous preacher had been 'driven half mad' by 'moral scavengers' who were 'daily darkening' the skies with 'new clouds of filth', and who deserved 'condign chastisement' for their 'wild, irresponsible' accusations (Hurlbert, 1874). In 1888 Hurlbert published a long account of his travels through Ireland. Its word usage echoes *Public Man*: 'quaint', 'mischief', 'disagreeable', 'absurd', 'ascertained', 'wildest', 'peremptorily', 'disquieted', and 'intimated'. Hurlbert's alliterations persisted: 'friendly financiers', 'rattled rapidly', 'exultingly exclaimed', 'utterly unlike' and so on (Hurlbert, 1888).

Hurlbert's writing style included many elements—an inventory of distinctive words, a fondness for vigourous verbs and certain '-ly' adverbs, a propensity for alliteration, and an uncanny ability to draw in the reader. Together these features provide the basis for 'traditional' attribution. None of the component elements alone would point to Hurlbert, but when combined together they are mutually reinforcing. Suddenly, the reader hears Hurlbert's unique voice. Traditional attribution, in the end, is more an art than a science.

Some of the contents of the Diary also point to Hurlbert rather than Ward. Twice the diarist mentioned Josiah Quincy (1772–1864), the retired President of Harvard College. Quincy was a living link between the New England Federalists who complained that the Louisiana Purchase opened the door to the expansion of slavery, and the New Englanders of the 1850s who reacted viscerally against the Fugitive Slave Law, the Kansas–Nebraska Act, and the resultant spectacle of 'Bleeding Kansas'. During the mid-1850s, when Hurlbert's own outlook on public affairs became most pro-Northern, he held the elderly ex-Federalist in high regard (Anonymous, 1879, entries for 28 Dec 1860, 4 Mar 1861; Hurlbert, 1856, p. 561, 568–69, 574–76).

The *Public Man* referred to others who would have been more familiar to Hurlbert than to Ward. During 1859 and 1860, Stephen A. Douglas enlisted several people to promote his presidential candidacy. Among them was George N. Sanders, a Kentucky native who was energetic, enthusiastic—and uncontrollable. Hurlbert and Sanders collaborated unsuccessfully to make the *New York Times* a Douglas paper. Following the election, however, the unpredictable Sanders became 'a loud and noisy secessionist', thereby earning a dismissive barb in the Diary. Far more than Ward, Hurlbert had reason to notice—and to ridicule—Sanders' erratic course (Anonymous, 1879, entry for 28 Feb 1861).

The diarist had a higher estimate of another insider from the Douglas campaign. John Forsyth, editor of the *Mobile Register*, was among the most outspoken champions of Douglas in the Deep South. The diarist claimed to have received a letter from Forsyth, before his arrival at the capital. The Forsyth reference, like the Sanders one, points to Hurlbert rather than Ward. Hurlbert hobnobbed with those who were close to Douglas; Ward did not (Anonymous, 1879, entry for 4 Mar 1861). One episode in the Diary reads very much as if it is about Ward, rather than by him. On 3 March the diarist encountered an 'anxious' visitor who had been enlisted by Seward to telegraph Jefferson Davis predicting that Lincoln's inaugural message 'would be conciliatory'. External evidence suggests that this was Ward (Anonymous, 1879, entry for 3 Mar 1861).

Anderson thought that the Diary ended abruptly on 15 March 1861 because that was when Ward must have left Washington, but here Anderson tried to have it both ways: he wanted the Diary that was not a diary to reflect Ward's movements (Anderson, 1948, p. 162). There are, however, even more powerful reasons why the Diary's author needed to end the document when he did. By 15 March it was generally assumed that Fort Sumter in Charleston Harbour soon would be relinquished. A few days after, though, doubts and questions began to arise. Perhaps the assumed decision had not been made after all? Perhaps the sigh of relief breathed by all who hoped for a peaceful resolution to the crisis was premature? Hurlbert, the dramatic stylist, initiated the Diary with the late-December flurry about Major Robert Anderson's move of his garrison from Fort Moultrie on Sullivan's Island to Fort Sumter. He ended the Diary at just the point when the Sumter difficulty supposedly had been resolved (Crofts, 1989, p. 289–90).

Anderson's judgment about the reliability of the Diary also must be qualified. It is now apparent that the Diary was created from scratch by a person who could not have been a legitimate diarist. It is, in fact, a memoir rather than a genuine diary. It was composed after the fact, not in 1861. The alleged diarist was a fictional construct—no such person ever existed. The Diary nonetheless was rooted in reality. Plenty of its inside information meets the historical test. Hurlbert wrote confidently about many matters that had remained secret at the time the Diary appeared, and that have since been authenticated. The Diary therefore must be taken seriously despite its seeming disqualifications (Crofts, 2010, p. 1–5).

Traditional attribution shows that the Diary's words reflect specific choices and patterns that were characteristic of Hurlbert's writing and entirely unlike Ward's. For an alternative and objective statistical analysis, we turned to the science of stylometry.

3 'Non-traditional' Attribution: Stylometry

3.1 Sampling and textual preparation

The stylometric task facing us was to examine the Diary and attribute it. As we have seen above, the contending authors are Samuel Ward and William Henry Hurlbert but two more possible authors were thrown into the stylometric mix: Henry Adams (1838–1918), the grandson and great-grandson of John Quincy and John Adams, and suggested as the diarist by Benjamin Price (1955); and James E. Harvey, one of the elite journalists of the antebellum era, who wrote for several different newspapers, including the *Philadelphia North American* and the *New York Tribune*. Other contenders—Allen Thorndike Rice, Thurlow Weed, James C. Welling, Richard Grant White and John W. Forney—were also investigated earlier in this research program, tested using computational methods, but were quickly rejected for statistical reasons as being unlike the Diary. These early contenders will not be discussed here.

A number of control texts are also necessary, in the same genre and preferably of the same era. Ideally, controls should be contemporary writers who do not feature as contenders to the disputed work. Accordingly, textual material from the letters and diaries of the following three people was put into machine-readable form for the analysis:

(1) Salmon Portland Chase (1808–73), who became one of the first Republicans elected to high office as governor of Ohio in 1855.

He failed to win the Republican presidential nomination in 1860, but was soon appointed Lincoln's Secretary of the Treasury (1861–64) and then as Chief Justice of the US Supreme Court (1864–73).

(2) George Templeton Strong (1820–75), whose voluminous diary published in four volumes in 1952 is full of pithy opinions and insights, and is the key to his posthumous reputation. He admired Lincoln, supported the Union war effort, and served on the US Sanitary Commission.

(3) Gideon Welles (1802–78), founder and editor of the *Hartford Times*. He was appointed Secretary of the Navy in 1861 by Lincoln and served until 1869. During his tenure in that office he kept a diary, which was published in three volumes in 1911.

Three textual samples, each of about 3,000 words, were taken from volume 1 of *The Salmon P. Chase Papers 1829-1872* (1993), from *The Diary of George Templeton Strong* and from volume 1 of the *Diary of Gideon Welles*.

For the two 'long-shot' contenders, five 3,000-word samples were taken from the volume *The Letters of Henry Adams 1858-1868* (1983) and three smaller samples from James Harvey's articles in the *Philadelphia North American and United States Gazette*, all written in 1860 and 1861. For our two main candidates the choices of source material were clear, given that we should control for genre. Three 3,000-word samples were taken from *Sam Ward in the Gold Rush* (1949) and three similarly sized textual samples from Hurlbert's travelogue *Ireland Under Coercion: The Diary of an American* (1888). Finally, for *The Diary of a Public Man* itself, four textual samples each of approximately 3,000 words, representing in total about two-fifth of the work, were taken at various places throughout the diary, being sufficiently spaced to enable a valid check to be made on internal consistency.

All these samples are listed in Table 1, the samples being either typed or scanned into machine-readable form. The choice of text size in stylometric studies is always problematic. Smaller text units are too short to provide opportunities for stylistic habits to operate on the arrangement of internal constituents, while larger units are insufficiently frequent to provide enough examples for reliable statistical inference. In their study of the *Book of Mormon*, Jockers *et al.* (2008) claim that even the smallest chapters are of adequate size for stylometric analysis, finding no correlation between the correct assignment of an author and the length of text sample. Forsyth and Holmes (1996) found the median text block size in a selection of stylometric studies to be around 3,500 words. The target length for this investigation was accordingly set at 3,000 words per sample but in the case of James Harvey we could only collect 1,400 words per sample. In all the analyses, the occurrence rates of words are measured in 'rates per thousand'. Thus for this study, differences in sample sizes are not critical provided we adhere to a stylometrically desirable minimum threshold of 1,000 words.

3.2 Stylometric methodology

In a pioneering work first published in 1964, Mosteller and Wallace employed frequencies of function words such as prepositions, conjunctions and articles as discriminators to investigate the mystery of the authorship of the *Federalist Papers*. Their scholarly analysis opened the way to the modern, computerized age of stylometry. The use of non-contextual high-frequency functions words as tools in attributional problems was continued by J. F. Burrows (1992) and since then multivariate statistical analyses involving large sets (50–100) of such words have met with astonishing success. Some noteworthy examples in a wide array of authors and genres have been the attribution of the 1583 *Consolatio*, shown to be not a lost work of Cicero but a 16th-century forgery (Forsyth *et al.*, 1999), the investigation into the authorship of the so-called 'Pickett Letters' of the American Civil War (Holmes *et al.*, 2001b), and the new look at the authorship of the *Book of Mormon* (Jockers *et al.*, 2008). The 'Burrows' approach essentially picks the N most common words in the corpus under investigation and computes the occurrence rate of these N words in each text or text-unit, thus converting each text into an N-dimensional array of numbers. Multivariate statistical techniques, most commonly

Table 1 Textual samples

Author	Title	Sample	*N* (sample size in words)
Salmon Chase	*Papers*	1	3,020
		2	3,000
		3	3,000
George Templeton Strong	*Diary*	1	3,130
		2	3,051
		3	3,056
Gideon Welles	*Diary*	1	3,102
		2	3,389
		3	3,070
Henry Adams	*Letters*	1	3,061
		2	3,007
		3	3,029
		4	3,038
		5	3,158
James Harvey	*Newspaper Articles*	1	1,396
		2	1,406
		3	1,375
Sam Ward	*Sam Ward in the Gold Rush*	1	3,016
		2	3,062
		3	3,043
William Hurlbert	*Ireland Under Coercion*	1	3,005
		2	3,004
		3	3,002
The Diary of a Public Man		1	3,222
		2	3,005
		3	3,058
		4	3,070

principal components analysis and cluster analysis, are then applied to the data to look for patterns. The former aims to reduce the dimensionality of the problem by transforming the *N* variables to a smaller number (usually 2) of new variables and the latter technique provides an independent and objective view of any groupings amongst the textual samples by means of a tree-diagram or dendrogram. The 'Burrows' approach has become the first port-of-call for attributional problems and will be the initial technique adopted in this investigation.

The value of *N* used varies by application and genre but typically lies between 50 and 75, the implication being that these words should be among the most common in the language and that content words should be avoided. Attributional studies have achieved success with *N* set at 50 (Holmes and Forsyth, 1995; Holmes, Robertson and Paez, 2001a), so this value of *N* is used as a rule-of-thumb heuristic throughout the analysis, being an appropriate value for these sized text samples. Future work with this textual data set could, however, include a study of the effect of varying the value of *N*. Appendix A lists these fifty most common function words, taken from the corpus of texts in Table 1.

In 2002 Burrows published a new way of using relative frequencies of common words for authorship attribution, naming his procedure 'Delta'. Burrows defines Delta to be 'the mean of the absolute differences between the z-scores for a set of word-variables in a given text group and the z-scores for the same set of word-variables in a target text'. The primary text with the smallest value of Delta, i.e. the smallest mean difference from the test text, is 'least unlike' it and the author of that primary text has the best claim, among the authors tested, to be the author of the

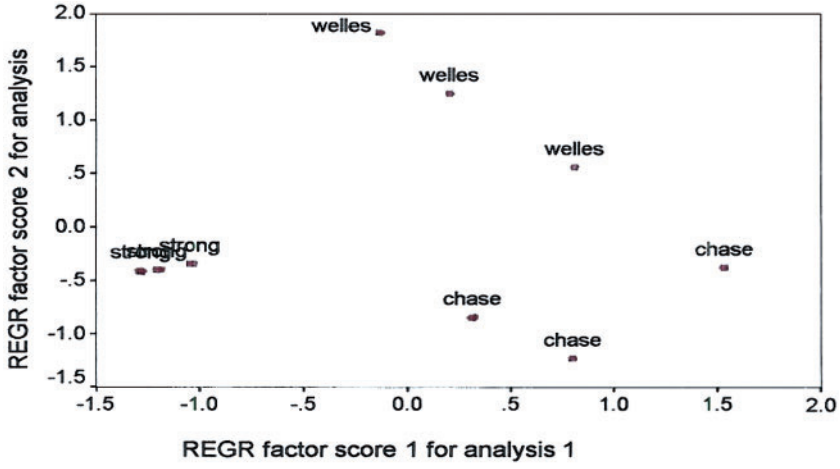

Fig. 1 Principal components plot: Chase, Strong, and Welles

test text. Delta has been subjected to extensive testing by Hoover (2004a) and was used by Jockers *et al.* in their study of Mormon scripture. Delta is becoming an important tool for stylometrists and is used in this study.

In a separate paper, Hoover (2004b) looked at alternatives or variants to Delta and showed that several variants were equally as effective as the original Delta in non-traditional attribution. His excellent results with these variants encouraged us to use them, also, in this investigation.

3.3 Hierarchy of analyses

3.3.1 *Controls: Chase, Strong, and Welles*

The first phase in this investigation was designed to test the validity of the proposed technique. For the purposes of this study, it is required that known texts can be shown to be internally consistent and separate from each other. The occurrence rates of the fifty words listed in Appendix A were computed for the individual textual samples from the diaries and papers of Chase, Strong, and Welles and used as input to both a principal components analysis and a cluster analysis. The positions of the samples in the space of the first two principal components, which together explain 55.1% of the variation in the original data, are shown in Fig. 1. An alternative analysis of the controls may be provided by conducting a cluster analysis on the textual samples, using the fifty word rates as variables and average linkage as the clustering algorithm. Figure 2 shows the resulting dendrogram.

Our results with these two methods of analysis are mutually supportive, with samples forming clusters on the basis of authorship. Our writers are internally consistent as regards their usage of these fifty words, yet are distinguishable from each other.

3.3.2 *Controls: Adams, Harvey, and Ward*

We now turn to three candidate authors of the Diary, albeit still in the 'control' mode of testing the validity of the 'Burrows' approach on textual samples from known writers. This time we will consider Samuel Ward, Henry Adams, and James Harvey. The occurrence rates of the 50 most frequently occurring function words were once again used as input to both a principal components analysis and a cluster analysis. The positions of the samples in the space of the first two principal components, which together explain 63.9% of the variation in the original data, are shown in Fig. 3. An alternative analysis of the controls was provided by conducting a cluster analysis on the textual samples, using the fifty word rates as variables. Figure 4 shows the resulting dendrogram.

Both our plots show quite remarkable internal consistency along with clear differences between Adams, Harvey and Ward on the basis of their use

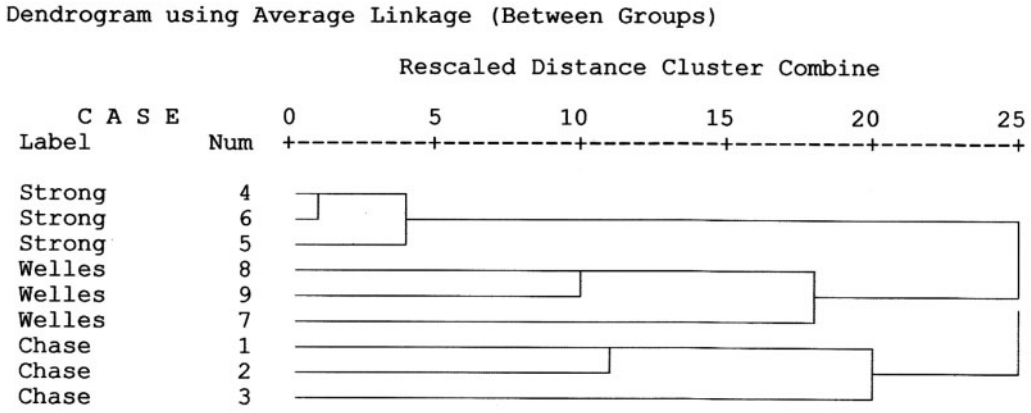

Fig. 2 Dendrogram: Chase, Strong, and Welles

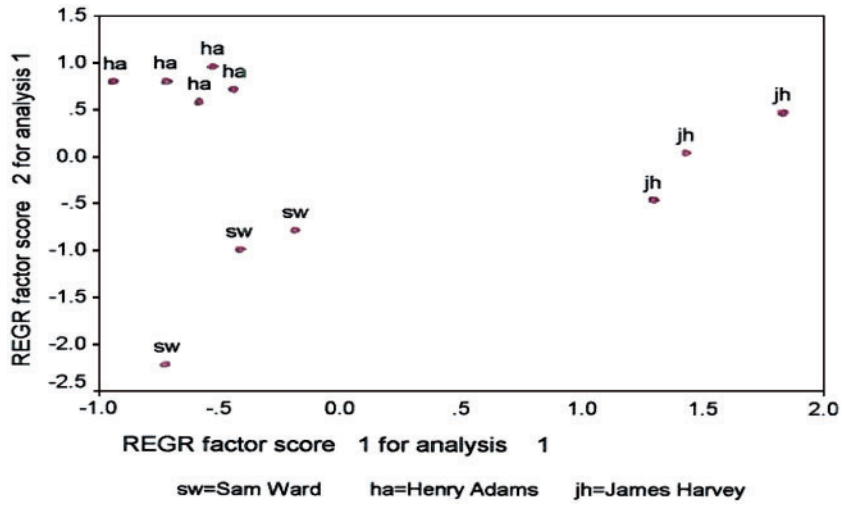

Fig. 3 Principal components Plot: Adams, Harvey, and Ward

of the function words. Once more the 'Burrows' approach has been validated on works of known authors.

3.3.3 *Adams, Harvey, Ward, and the Diary*

Having successfully validated the technique on the control samples, we may now incorporate the textual samples from the Diary into the analysis. Accordingly the next step was to add the Diary samples to the three candidate authors in (b) above. Once more the occurrence rates of the fifty most frequently occurring function words were used as input to both a principal components analysis and a cluster analysis. The positions of the samples in the space of the first two principal components, which together explain 52.7% of the variation in the original data, are shown in Fig. 5. The alternative analysis provided by conducting a cluster analysis on the textual samples, using the 50 word rates as variables, revealed the dendrogram shown in Fig. 6.

Two important conclusions may be drawn from these clear and mutually supportive plots. First, the four Diary samples show excellent internal consistency, suggesting single authorship. Secondly, the Diary samples cluster quite distinctly and separately from the samples of the writings of Adams, Harvey,

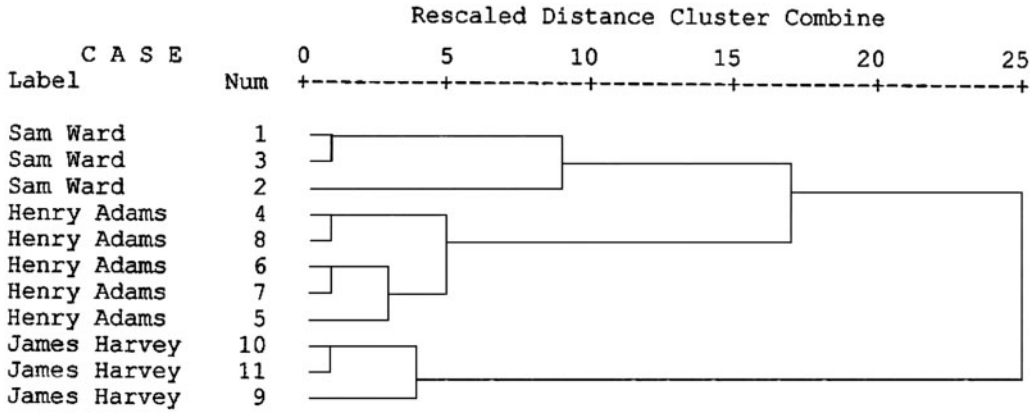

Fig. 4 Dendrogram: Adams, Harvey, and Ward

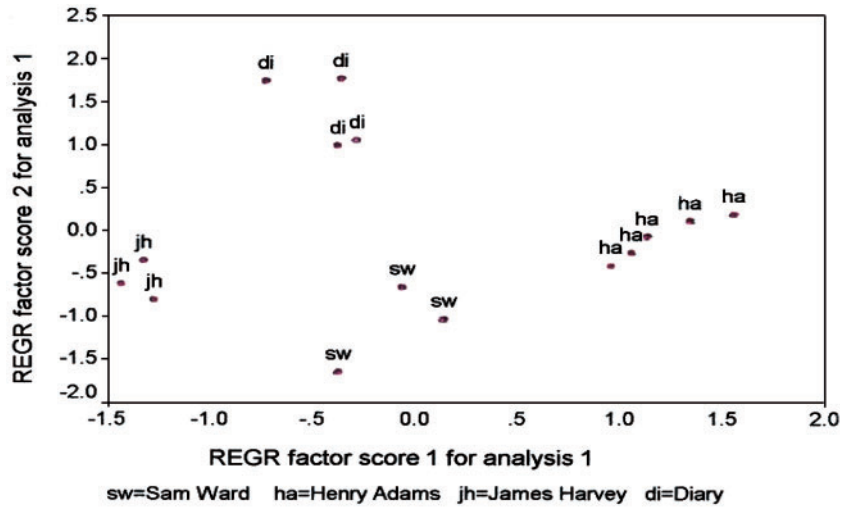

Fig. 5 Principal components plot: Adams, Harvey, Ward, and the Diary

and Ward. Two of our possible contenders and one of our strong contenders for authorship all now appear not to match the author of the Diary according to the 'Burrows' approach.

3.3.4 *Ward and Hurlbert*

The control samples have served their purpose and it is now time to bring the Hurlbert samples into the analysis. Leaving aside the Diary samples for the moment, it will be interesting simply to match the two main contenders Ward and Hurlbert against each other. Using the 'Burrows' approach as in the above analyses, Figs 7 and 8 show the principal components plot and the dendrogram for the textual samples from these writers. Both these supporting plots show that Ward's and Hurlbert's function words differ in their rate of usage.

We may explore this separation by looking at Fig. 9, the associated scaled loadings plot from the principal components analysis, which helps to explain the groupings in the main plot. One can imagine superimposing this graph on top of the principal components plot. Words on the right such as 'no', 'is' and 'be' have high usage by

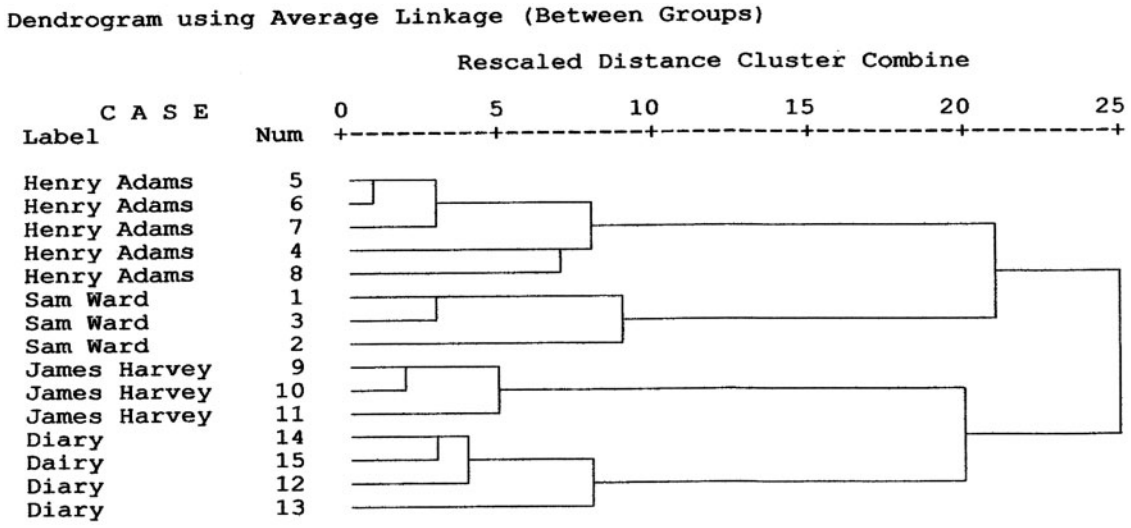

Fig. 6 Dendrogram: Adams, Harvey, Ward, and the Diary

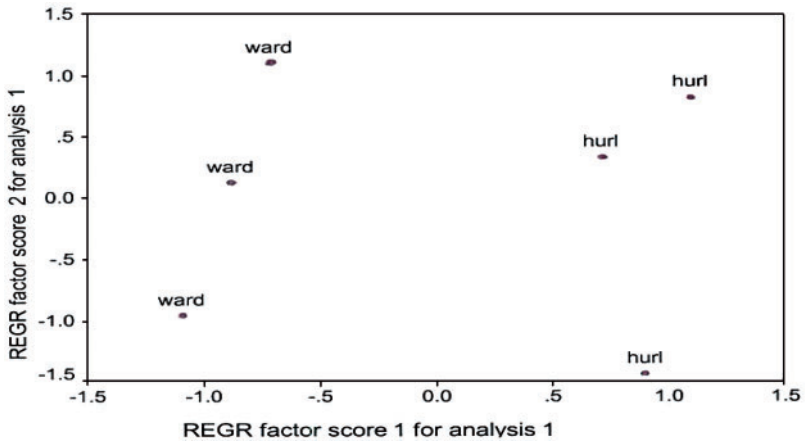

Fig. 7 Principal components plot: Ward and Hurlbert

Hurlbert, whereas words on the left such as 'with', 'which' and 'had' are words favoured by Ward.

Following the results from all our analyses, we may now confidently proceed to the final stage of the non-traditional attribution, namely comparing the Diary solely with Ward and Hurlbert.

3.3.5 *Attribution: Ward, Hurlbert and the Diary*

For the attributional phase of this investigation we turn to a third technique in multivariate statistical analysis, namely discriminant analysis. Essentially, this technique shows how well it is possible to separate two or more known groups of cases (samples), given measurements on several variables for these samples. The technique uses the known data to derive one or more 'discriminant functions'. These functions are linear combinations of the variables that best separate the known groups. The functions are then tested against the known data to assess the classification accuracy of the routine. Once the functions have been tested, they may then be used

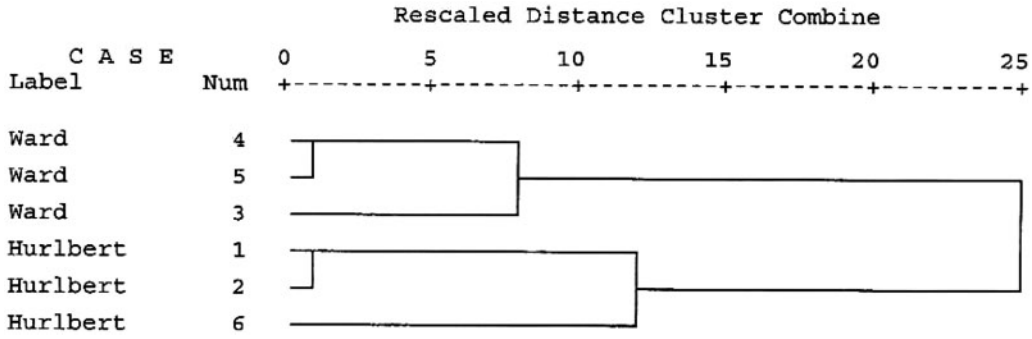

Fig. 8 Dendrogram: Ward and Hurlbert

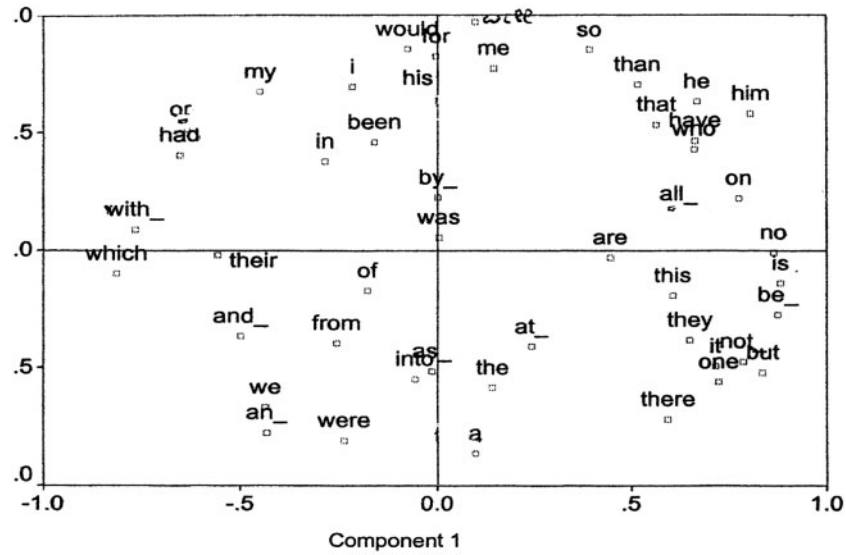

Fig. 9 Loadings plot: Ward and Hurlbert

to classify cases for which the true group is not known, the cases being assigned to the group that they are 'closest' to in a multivariate sense.

Prior to the use of discriminant analysis, however, it would be instructive to conduct another principal components analysis using our 50 words and view the positions of the Ward, Hurlbert and Diary samples in the space of the first two components. Figure 10 shows this plot .

Projection onto the first, dominant, component in Fig. 10 hints that Hurlbert may be 'closer' to the Diary than Ward but we should note the low percentage of variation explained (53.5%) by looking at just two dimensions.

To use discriminant analysis, it is advisable not to have more variables than cases. The Ward samples were combined and then split into nine new samples, each of approximately 1,000 words. Similarly the Hurlbert samples were combined and then subdivided into nine new samples, each of approximately 1,000 words. We thus have a total of 18 new samples, each of approximately the same size, drawn from two known and distinctive groups. The Diary samples were next combined

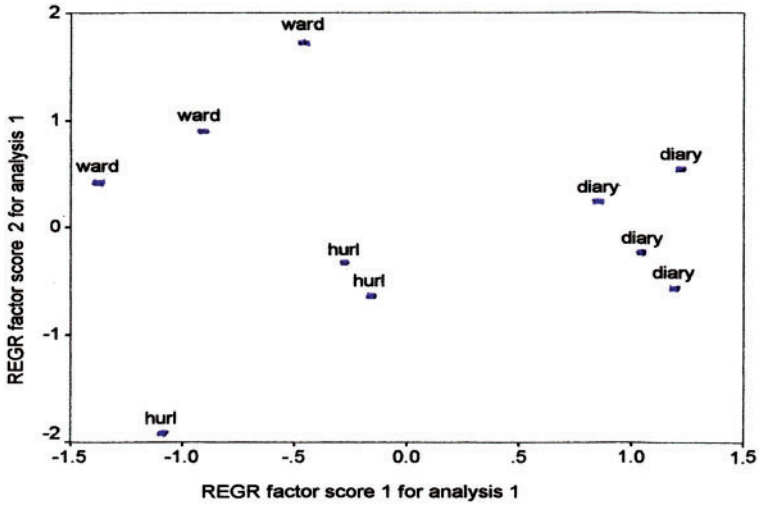

Fig. 10 Principal components plot: Ward, Hurlbert, and the Diary

and subdivided into 12 new samples, each, once more, of approximately 1,000 words. To run the discriminant analysis procedure, the number of variables was reduced to the thirty most frequently occurring function words in the Ward–Hurlbert–Diary corpus. These words are listed in Appendix B.

An initial, exploratory, look at the groupings of our new samples may be obtained by looking at the principal components plot of our new (30×30) data matrix. Figure 11 shows this plot.

This plot, again, suggests that Hurlbert lies 'closer' to the Diary than Ward, but we need to turn to discriminant analysis in an attempt to resolve this question.

For a two-known-group analysis, one discriminant function is derived. Table 2 gives information on the effectiveness of the discriminant analysis procedure on our known data. A test of the null hypothesis that, in the populations from which the samples are drawn, there is no difference between the two group means is based on the Wilks' Lambda statistic. This hypothesis may be firmly rejected ($P < 0.0005$) implying that the two style groups, Ward and Hurlbert, are significantly different with respect to the means of their discriminant scores.

Table 3 shows the 11 words which load most heavily on the discriminant function.

It is instructive to now compare this table with the loadings plot for Ward and Hurlbert shown in Fig. 9. Highly positively loaded words from the discriminant function such as 'be', 'is', 'have' and 'it' are all words which appear on the right (Hurlbert) side of the loadings plot in Fig. 9, while highly negatively loaded words such as 'which', 'with' and 'had' are words which appear on the left (Ward) side of the loadings plot.

Another measure of the effectiveness of the classification routine is seen in the top half of the SPSS classification matrix in Table 4. All 18 of our known textual samples have been correctly classified into their style groups, even with cross-validation—a more rigorous classification algorithm where each sample is, in turn, omitted, the discriminant function derived on the 17 remaining samples, then the omitted sample allocated to its nearest style group.

The classification routine was then instructed to assign the 12 samples from the Diary to the group (style) they are closest to, in the sense of distance in multivariate space. The lower half of the classification matrix in Table 4 shows that all 12 of the Diary samples have been placed in the Hurlbert style group.

3.4 'Delta' analysis

For an alternative non-traditional stylometric investigation into the authorship of the Diary, we turn to Burrows' Delta technique and the variants to Delta introduced by Hoover, first discussed in section 3.2

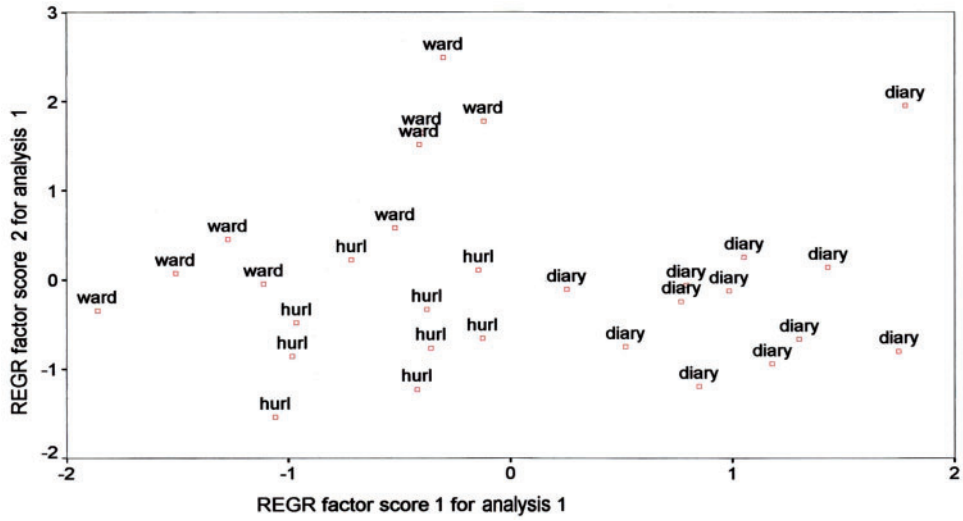

Fig. 11 Principal components plot: Ward, Hurlbert, and the Diary using new samples

Table 2 Wilks' lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	0.001	71.642	10	0.000

Table 3 Standardized canonical discriminant function coefficients

	Function 1
had	−1.489
as	1.502
it	10.001
be	26.203
which	−11.030
that	2.967
on	6.614
is	25.529
with_	−6.987
this	4.078
have	16.445

above, and labelled Delta Prime. For this analysis, four potential authors of the Diary are measured: Ward, Adams, Harvey, and Hurlbert, and the primary samples formed by combining the samples used for the earlier tests. Both Burrows (2002) and Hoover (2004a) found that the accuracy of Delta increases as the number of frequent words included in an analysis rises from 40 to 150. Jockers *et al.*, (2008) use 110 words in their Delta tests. Initially we set the number of words at 50 but following private correspondence with Burrows, it was decided to use the top 100 most frequently occurring function words for this analysis. Appendix A lists the additional fifty words now added to the top fifty used in the earlier analyses involving principal components and clustering.

We now add a second corpus from William Hurlbert. In addition to his travelogue *Ireland Under Coercion: The Diary of an American*, now labelled 'Hurlbert 1', Hurlbert wrote lengthy articles for the *New York Times* which were published in September and October 1862. These articles total 16,451 words and were added to the text corpus used for Delta and Delta Prime analysis under the label 'Hurlbert 2'.

Table 5 below lists the Delta and Delta Prime indices of 'closeness' between the Diary and the candidate authors:

Values of Delta typically range from 0.5 to 1.5 (Hoover, 2004a). From this table we can see that the lowest Delta score (0.7784) is that of Hurlbert; in other words Hurlbert is 'least unlike' the Diary author and has the best claim to be the true

Table 4 Classification results

			Author	Predicted Group Membership		Total
				Ward	Hurlbert	
Cases Selected	Original	Count	Ward	9	0	9
			Hurlbert	0	9	9
		%	Ward	100.0	0.0	100.0
			Hurlbert	0.0	100.0	100.0
	Cross-validated	Count	Ward	9	0	9
			Hurlbert	0	9	9
		%	Ward	100.0	0.0	100.0
			Hurlbert	0.0	100.0	100.0
Cases Not Selected	Original	Count	Ward	0	0	0
			Hurlbert	0	0	0
			Ungrouped Cases	0	12	12
		%	Ward	0.0	0.0	100.0
			Hurlbert	0.0	0.0	100.0
			Ungrouped Cases	0.0	100.0	100.0

Table 5 Delta and delta prime scores

Evaluation methods	**Evaluation of authorship of the diary of a public man**				
	The diary of a public man versus Sam Ward	**The diary of a public man versus Henry Adams**	**The diary of a public man versus Harvey**	**The diary of a public man versus Hurlbert 1**	**The diary of a public man versus Hurlbert 2**
Mean delta	**1.0115**	0.9154	0.8415	*0.7784*	0.7854
Delta prime: square of positive mean difference minus negative mean difference	**2.0507**	1.7331	1.5556	*1.3847*	1.3847
Delta prime: twice positive mean difference minus negative mean difference	**3.0400**	2.7204	2.5388	*2.3329*	2.3692
Delta prime: three times positive mean difference minus negative mean difference	**4.1434**	3.6077	3.4090	*3.1054*	3.2448
Delta Prime: square of (positive mean difference plus one) minus negative mean difference	**5.2575**	4.5077	4.2961	*3.9296*	4.1358

Bold value indicates maximum value. Italic value indicates minimun value.

author from our four contenders. Importantly, the second lowest Delta score (0.7854) is also from Hurlbert—his *New York Times* articles. Thus, both the text corpora from Hurlbert match the Diary better than text from Ward, Adams, and Harvey.

The first variant of Delta, labelled Delta Prime and introduced in detail by Hoover (2004b), also suggests Hurlbert as author, with the lowest Delta Prime score (1.3847) obtained on both Hurlbert corpora. Hoover notes that this particular re-definition of Delta produces results slightly more accurate than those based on Delta itself. The remaining three variants are simply transformations of Delta Prime that weigh positive differences

more heavily than negative ones. Although there is no theoretical basis for these transformations, and one must expect them to be correlated, Hoover presents a powerful inductive basis for their employment. Both the Hurlbert 1 and Hurlbert 2 corpora are of closer match to the Diary using Delta Prime and its three transformations than Ward, Adams and Harvey.

These remarkable and conclusive findings, coupled with the evidence from the hierarchical analyses in section 3.3 above, confirm that non-traditional authorship attribution points very strongly to William Hurlbert being the author of the Diary.

4 Conclusion

The non-traditional analysis has supplied objective, stylometric evidence that supports the traditional scholarship on the authorship of the Diary. While, in the absence of definitive external evidence, no attributional claim can be absolute, some methodologies will nevertheless be more reliable than others. In blending a traditional approach to the attribution of *The Diary of a Public Man* with a non-traditional stylometric approach, we agree with the viewpoint of Hänlein (1999), who argues that the most reliable results in authorship recognition studies take into account both 'intuitive' findings—i.e. the traditional scholar's inherently subjective recognition of an author's distinctive style—and computational methods. A sequential approach to attribution is also recommended by Rudman (1998), who stresses: 'Any non-traditional study should only be undertaken after an exhaustive traditional study. The non-traditional is a tool for the traditional authorship scholar, not a proving ground for statisticians and others to test statistical techniques'.

This joint interdisciplinary approach to a problem that has fascinated historians for over a century concludes that William Hurlbert has the strongest claim to be the true author of the Diary. We recommend this stepwise procedure to other researchers studying attribution problems.

Acknowledgements

We wish to thank The College of New Jersey for its support during this research program, and former students Carol Antenna, Kelliann Brennan, Ryan Christiansen, John Rutledge and Satwinder Thind for their invaluable assistance in textual acquisition and preparation. We also wish to thank Dr Richard Forsyth of the University of Southampton, UK, for the specialist computer software and Dr David Hoover of New York University for his assistance with Delta Prime.

References

Anderson, F.M. (1948). *The Mystery of "A Public Man": A Historical Detective Story*. Minneapolis: University of Minnesota Press.

Anonymous, (1879). The diary of a public man: unpublished passages of the secret history of the American Civil War. *North American Review*, **129:**(Aug. 1879) 125–40; (Sept. 1879) 259–73; (Oct. 1879) 375–88; (Nov.1879) 484–96.

Bancroft, F. (1896). Letter to James Ford Rhodes. 19 Aug. *James Ford Rhodes Papers.* Massachusetts Historical Society.

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

Burrows, J.F. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, **17**: 267–87.

Collins, C. (ed.) (1949). *Sam Ward in the Gold Rush*. Stanford: Stanford University Press.

Crofts, D.W. (ed.) (1989). *Reluctant Confederates: Upper South Unionists in the Secession Crisis*. Chapel Hill: University of North Carolina Press.

Crofts, D.W. (2010). *A Secession Crisis Enigma: William Henry Hurlbert and 'The Diary of a Public Man'*. Baton Rouge: Louisiana State University Press. [The full Diary appears as an appendix to this volume].

Forsyth, R.S. and Holmes, D.I. (1996). Feature-finding for text classification. *Literary and Linguistic Computing*, **11**: 163–74.

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

Hänlein, H. (1999). *Studies in Authorship Recognition – A Corpus*-based *Approach*. European University Studies, Series XIV, Vol. 352. Frankfurt am Main: Peter Lang.

Harvey, J. (1860–61).Reports from Washington for the *Philadelphia North American and United States Gazette*, 23, 24, 26 December 1860, 30, 31 January, 1, 3, 5, 8, 9 February 1861.

Holmes, D.I. and Forsyth, R.S. (1995). The 'Federalist' revisited: new directions in authorship attribution. *Literary and Linguistic Computing*, **10**: 111–27.

Holmes, D.I., Robertson, M., and Paez, R. (2001a). Stephen Crane and the 'New-York Tribune': a case study in traditional and non-traditional authorship attribution. *Computers and the Humanities*, **35**: 315–31.

Holmes, D.I., Gordon, L.J., and Wilson, C. (2001b). A widow and her soldier: stylometry and the American Civil War. *Literary and Linguistic Computing*, **16**: 403–20.

Hoover, D.L. (2004a). Testing Burrows' delta. *Literary and Linguistic Computing*, **19**: 453–75.

Hoover, D.L. (2004b). Delta prime?. *Literary and Linguistic Computing*, **19**: 477–95.

Hurlbert, W.H. (1854). *Gan-Eden: or, Pictures of Cuba*. Boston: J.P. Jewett & Co.

Hurlbert, W.H. (1856). The political crisis in the United States. *Edinburgh Review*, **104**(1856)561–97.

Hurlbert, W.H. (1859). *New York Times*editorials dated 19, 27 October, 2, 3, 4, 5, 14, 17, 21, 23, 25, 26 November, 1, 3, 5, 6, 7, 12, 13 December 1859.

Hurlbert, W.H. (1862). *New York Times*. articles dated 10, 11, 15, 23 September, 4, 11, 20, 30 October 1862.

Hurlbert, W.H. (1868). *New York World*. editorials dated 24, 25, 26 February, 2 March 1868.

Hurlbert, W.H. (1874). *New York World*. editorials dated 27 July, 1, 13 August 1874.

Hurlbert, W.H. (1888). *Ireland Under Coercion: The Diary of an American*. Edinburgh: David Douglas.

Jockers, M.L., Witten, D.M., and Criddle, C.S. (2008). Reassessing authorship of the 'Book of Mormon' using delta and nearest shrunken centroid classification. *Literary and Linguistic Computing*, **23**: 465–91.

Levenson, J. C. (ed.) (1983). *The Letters of Henry Adams, Vol. 1*. Harvard: Harvard University Press.

Mosteller, F. and Wallace, D.L. (1964). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Reading. MA: Addison-Wesley.

Nevins, A. (ed.) (1952). *The Diary of George Templeton Strong*. Macmillan.

Niven, J. (ed.) (1993). *The Salmon P. Chase Papers, Vol. 1, 1829-1872*. Kent: The Kent State University Press.

Potter, D.M. (1942). *Lincoln and His Party in the Secession Crisis*. New Haven: Yale University Press.

Rudman, J. (1998). Non-traditional authorship attribution studies in the 'Historia Augusta': Some Caveats. *Literary and Linguistic Computing*, **13**: 151–7.

Welles, G. (1911). *Diary of Gideon Welles,* Vol.1. New York: Houghton Mifflin.

Appendix A One hundred most frequently occurring function words

Top fifty words used in the principal components and cluster analyses, and also in the Delta analysis.				
the	is	but	you	will
of	as	at	him	an
and	was	by	all	no
to	for	this	would	they
a	be	on	are	were
in	with	from	if	one
i	his	had	been	we
that	not	my	has	there
he	have	me	who	their
it	which	or	so	than
Next fifty words, used in the Delta analysis only.				
more	should	time	these	like
any	some	out	after	little
them	now	about	see	men
do	much	your	general	life
when	may	am	upon	most
what	said	other	then	before
our	new	such	two	day
very	here	great	made	up
can	could	into	must	say
its	only	her	good	last

Appendix B Thirty words used in the discriminant analysis

the	with	be
of	as	on
to	it	this
and	his	me
a	for	is
in	which	from
he	at	him
i	had	my
that	not	were
was	by	have