Attribution by Statistics : A Critique of Four Recent Studies

M.W.A. SMITH

Abstract

Of a diverse range of statistical techniques available to scholars wishing to address problems of authorship, only a few have attracted critical attention. In some, odds of millions to one against error of attribution are calculated. It is argued that such long odds merely indicate an absence of scepticism on the part of some stylometricists. A recent study which has been both lauded unreservedly as a pioneering advance, and condemned outright as a misuse of a familiar technique, is then examined. Critiques of two other recent studies are included to illustrate the pitfalls which await the unwary scholar. Next, a theoretical basis is developed for a method which shows considerable promise. As a consequence, a departure from the theory is found and the necessity of rectifying the procedure is illustrated. Finally, prerequisites and principles for resolving problems of authorship are discussed.

An Examination of Virtual Certainty of Attribution

"Computer-assisted, statistically-based stylometry is a potentially valuable tool for use ... in authenticity studies ... It substitutes quantification ... for 'literary connoisseurship' ... The new stylometry ... requires the analyst to survey every word in the sample ... [so that] a sufficient number of writing habits can be identified to establish statistical validity. Tests consisting of twenty such habits reduce the probability of error to almost the vanishing point"[1]. Thus Harold Metz concluded a paper which is based on tables, prepared by A.Q. Morton, of frequently occurring

[1] G. Harold METZ : Disputed Shakespearean Texts and Stylometric Analysis. *Text : Transactions of the Society for Textual Scholarship*. 2(1985), 149-71; p.156.

words in certain plays by Shakespeare and Peele. Doubtless he was influenced by Morton who had written, " ... even if the comparison is restricted to the five tests... the probability that the works of Peele belong to the same population as the three plays of Shakespeare is less than one in ten thousand million"[2].

Similarly D.J. Lake found that his statistical investigation pin-points Middleton as the author of *The Revenger's Tragedy* (c. 1606) to a level of accuracy of 1 in 100 million or 1 in 15000 million (depending on the variant of the method chosen) of the population and that the figures are "almost certainly gross underestimates"[3].

Possibly the record for large numbers is held by Thomas Merriam. In response to Eric Sams's claim that the anonymous play *Edmund Ironside* is an early work by Shakespeare, he is reported to have concluded from statistical tests that Robert Greene was 890 million, million, million times more likely to be its author[4].

What do such figures actually mean? If, for instance, a person who drinks six cups of tea every day had the habit of spilling merely one in every million, on average he would do so only once every 456 years. Were he to improve so that the chance is decreased to one in 15000 million, such an accident would befall him on average only once in every 6.8 *million* years. Clearly in offering such figures for reliability in an application so beset by uncertainty, the investigators must have become mesmerised by their arithmetic at the expense of their critical faculties[5].

If studies which claim such enormous odds for the reliability of their conclusions so readily invite suspicion, what of those which more modestly offer no probabilities whatever? Of the four recent contributions examined in this paper, only the third introduces probabilities. The first has attracted unqualified praise from one quarter for its originality and for the advance it represents over earlier techniques while

---

[2] See note 1, p. 155.

[3] David J. LAKE : *The Canon of Thomas Middleton's Plays*. Cambridge University Press, London, 1975; p. 156.

[4] Nigel HAWKES : Play on words blanks out the Bard's drama. *The Observer*, 12 January 1986, p.3.

[5] For a critical examination of A.Q. Morton's main work on English texts, see :

(i) M.W.A. Smith : An Investigation of the Basis of Morton's Method for the Determination of Authorship. *Style* 19(1985), 341–68;

(ii) M.W.A. Smith : An Investigation of Morton's Method to Distinguish Elizabethan Playwrights. *Computers and the Humanities* 19(1985), 3–21 and 144;

(iii) M.W.A. Smith : Forensic Stylometry : A Theoretical Basis for Further Developments of Practical Methods. *The Journal of the Forensic Science Society* 29(1989), 15–33.

For an investigation of Lake's conclusions see :

(iv) M.W.A. Smith : *The Revenger's Tragedy* : The Derivation and Interpretation of Statistical Results for Resolving Disputed Authorship. *Computers and the Humanities* 21(1987), 21–55 and 267.

elsewhere being condemned as merely a misinterpretation of an old familiar formula. The second is the sort of application which could well perpetrate the simplistic impression that any differences, detected as a consequence of the execution of computer packages on a few texts, are sufficient to reveal authors hip and/or chronology. A critique of the procedures used exemplifies the pitfalls awaiting the unwary. In contrast, the next study describes a statistical method in which its authors are careful to examine alternative explanations for their findings; they thereby avoid jumping to misleading conclusions on the basis of impressive odds. Nevertheless, the importance of penetrating the details before full acceptance of any such experiment is illustrated by the questions which arise as to the degree of accuracy with which these investigations were actually performed. The fourth method has been presented as a practical procedure heavily depended on eigenvalues. Because it shows considerable potential, its theoretical foundation is deduced. As a consequence, a departure from the theory is found and the necessity of rectifying the procedure is clarified.

The paper ends with a general discussion of the prerequisites and principles for resolving problems of authorship.

A Rediscovery of Type-Token Ratios

In 1988 John Baker suggested that the rate at which an author introduced new words into his works should be called his *pace* and that this measure had generally been overlooked by researchers in the field of attribution by quantitative methods[6].

Darrel Ince, Professor of Computing at the Open University (England), declared that Baker's method "has at least three advantages over past work. First, it is computationally very simple and does not require very much programming or advanced statistics. Second, it can be applied to works of different genres such as essays, plays and novels. Third, it matches our perceptions of how great writers grow in stature as they get older"[7]. Ince informed his readers that this "recent pioneering work ... departs radically from previous research" and that Baker "has developed a measure which reflects an author's ability to generate new words as the length of a manuscript increases." He reported that the measure used "is the ratio of the total number of words in the vocabulary to the total number of words in the text. This measure, known as pace, measures the maturity of an author ... Baker's thesis ... is

[6] John Charles BAKER : Pace : A Test of Authorship Based on the Rate at which New Words Enter an Author's Text. *Literary and Linguistic Computing* 3(1988), 36-9.

[7] Darrel INCE : Authors put under computer scrutiny. *The Independent*, 14 August 1989, p.14. According to the *Commonwealth Universities Yearbook* 1989, Vol 1, p.746, Professor Ince's main interest is Mathematics.

that as an author ages, his pace increases." Ince believes that while the "technique has a major use as a tool for dating literary works, it also holds out the promise that the authorship of anonymous works can be determined ... The method is based on bisecting the anonymous work and combining it with a work whose attribution is certain. If the pace of the combined piece changes, then the author whose work has been used has not written the anonymous work. However, if it stays constant, then there is a high probability that the author has been found."

A far more realistic critique has been recorded by Louis Milic. He noted (while pointing out errors in Baker's presentation) that *pace* is simply type-token ratio (TTR) and therefore neither novel nor pioneering, that authors do not exhibit unique TTRs but instead reveal overlapping ranges and that TTR itself is actually very sensitive to the length of the text[8].

Milic's comments are flawed by only one lapse : he seemed to have overlooked that Baker had stated, "an author's rate of generating new words is reasonably independent of the word length of his manuscript - once the text has passed certain limits, i.e., an extremely short text will understandably rank remarkably rich and extraordinarily longer ones will appear more impoverished." Thus, as a correction of Baker, Milic's first table is irrelevant. On the other hand, had Baker himself attempted to establish values for these limits, he would doubtless have been forced to abandon any such notions of constancy.

Baker's hypothesis can be seen to be false without reference to external data. From his own figures for Sestiad 3 combined with Sestiad 4 and for Sestiad 5 with 6 of *Hero and Leander*, the TTR for each exceeds substantially that for Chapman's contribution as a whole, i.e. Sestiads 3 through 6. More interesting, however, is his statement, "Thomas Nashe writing at different times in three entirely different styles or genre, in lengths ranging from 11,500 words to 27,391 words generated types, new words, at precisely the same pace throughout : 23 % ..." Such an observation could equally be evidence for the opposite : if there were an increase in repetition of words with increasing length of text, then if TTRs of non-overlapping blocks of 11,500 consecutive tokens were taken from the longest piece, the average value would be expected significantly to exceed 0.23. If so, Baker's constant figure would actually accord with traditional wisdom that TTR decreases with increasing length of text and that authors exhibit considerable variability in this measure. Far from exemplifying a new insight, the most likely explanation is that both effects have combined to produce the same figure for all three texts.

---

[8] Louis T. MILIC : A comment on John Baker's article. *Literary and Linguistic Computing*, 4(1989), 153-4.

A simple but telling inference is that scholars should not accept on trust a paper published in a reputable journal even when it attracts unqualified support from an apparently reliable quarter : the carefully considered opinions of several capable critics should be sought.

Spanish Stylometry

Another contribution to the resolution of authorship may initially appear as if it could serve as a model for further studies. Its authors, Stratil and Oakley, also claimed "a new way of dating plays by Tirso de Molina"[9]. Their purpose was "to achieve a quantitative mathematical analysis and content analysis of three plays by Tirso de Molina." They reported that such an attribution is certain only in the case of *El vergonzoso en palacio*, probable for *El burlador de Seville* and "hotly disputed" for the third play *El condenado por desconfiado*. Therefore, the latter two were "the subject of an authorship analysis carried out at the University of Birmingham [England]." As a basis for comparison Act I of *El mágico prodigioso* by Calderón is included.

The first set of tests is based on sentences. These plays are Spanish and date from the 17th century. In English plays of the same period, texts frequently are corrupt and punctuation cannot be taken as authorial without thorough investigation. In a footnote, Stratil and Oakley mentioned corrupt texts of *Burlador*. It therefore seems possible that equivalent textual problems could bedevil the study of early Spanish plays. Accordingly, it would have been pertinent for the authors to have started by discussing the validity of comparisons which depend on the pointing. As it turns out, however, such preliminaries would not have affected the judgement on this part of their study; Stratil and Oakley merely provided a verbal description of the graphs they had obtained, claiming that complexity and dependency on colour prevented reproduction of their computer-produced figures. From their descriptions, all the graphs, in fact, would appear to be straightforward and should have been redrawn appropriately for publication. Moreover, the technique used to obtain the graphs should have been reported properly, not merely by mentioning an undefined "normalisation routine" and an unquoted (and unreferenced) "mathematical formula". In principle it is meaningless to report work without offering the reader the means to check. In this particular case it does not matter much because another deficiency is that the only basis for comparison is one play of undisputed authorship and one Act of another known to be by a different writer. Indeed the authors' dubious

[9] Marie STRATIL and R.J. OAKLEY : A Disputed Authorship Study of Two Plays Attributed to Tirso de Molina. *Literary and Linguistic Computing* 2(1987), 153-60.

interpretation of results is evident from the few figures presented in this section. Average sentence lengths for Act I of the plays *Vergonzoso, Burlador* and *Mágico* are given as 12.789, 10.029 and 14.479, respectively. Stratil and Oakley stated, "We see that Tirso's average sentence length is between two and four words shorter that Calderón's." With so little data to hand, their approach is to assume that *Burlador* is by Tirso and then, here as elsewhere in their paper, to look for similarities or differences as appropriate. Emphasising the importance of the principle of enabling the reader to check assertions made by the writer(s) of a paper, it can be seen, according to these figures, that another interpretation is that Calderón is closer to Tirso than Tirso is to the author of *Burlador.*

At one point in their paper Stratil and Oakley stated, "It would be desirable to have similar analyses of plays by different authors to decide to what extent these patterns represent an author's imprint on the language and to what extent they show a natural trend of the Spanish language." In fact, such comparisons are not merely desirable but absolutely essential. In every section of their paper conclusions or inferences are presented for which there is nothing even approaching acceptable evidence. For instance, by assuming implicitly that the attributions to Tirso are correct and then relying on existing theories, Stratil and Oakley suggested that differences they found in comparisons based on sentences are due to "a strange interference" in the text or to "the same writer at a different period." Similarly, a reasonable closeness of results for word-length in the third and seventh positions in sentences is interpreted, without comparison with plays by other contemporaneous writers, as indicating "almost certainly ... that the same person had written all three texts." Unfortunately the authors have neglected the fundamental prerequisite for any test, which is, to determine whether or not it has the capability to differentiate between writers.

The next section consists of a "Total Word Count Analysis" in which the term *distinct words* is used for *types, running words* for *tokens* and both *Ratio R/D* and *repetition ratio* for the reciprocal of the type-token ratio. The three plays attributed to Tirso are again tested without reference to other playwrights' dramas. The percentages of words of lengths 1 to 15 letters for each of the plays is calculated and for each Act the *Ratio R/D* is determined. Stratil and Oakley concluded that because "a consistent pattern between the three plays is again unmistakable ... these results may point to the same author ... [but] a disturbance is still apparent in the *Burlador* figures as they show the highest and lowest ratio of repetition." Neither of the criteria used in this section are sensitive to the authorship of English plays of the period but this does not necessarily mean that they should be discarded for use in other languages. Again, a programme of validation would have been

obligatory and, if successful, an investigation consisting of appropriate comparisons with plays by other authors should have been designed. Without evidence of authorial variability it is not possible to attribute 'disturbance' to 'interference', particularly when authorship itself is uncertain.

Stratil and Oakley next took the 30 most frequent words in each of the nine Acts of the three plays by, or attributed to, Tirso. Forty-four words were thereby obtained of which "just under half ... are represented among the most frequent [Juilland] dictionary words." Stratil and Oakley noted that the difference "in the age of the language" of the 55 plays used in the dictionary and the language of the 17th century plays in question could be a factor in such a difference of usage. Nevertheless, because they found this difference consistent and because 30 percent of the 44 words are excluded from the list in the dictionary of words occurring more than 15 times, they "deduce ... that we may well have a very individualised use of these words [and that] their homogeneity would point to the same writer for all three plays." This is yet another remarkable conclusion given the absence of comparative data. Moreover, such an assessment was underpinned presumably by merely inspecting visually a table of occurrences consisting of 44 rows and 9 columns of figures.

The numbers of tokens and types given by Stratil and Oakley for *Vergonzoso* are :

	Running words (i.e. tokens)	Distinct words (i.e. types)
Act 1	6445	1914
Act 2	6353	1754
Act 3	3978	1228
Total	16776	4896

The total for types is, of course, meaningless because the number of types in the entire play is doubtless less than 4896. More importantly they stated, "the level of running words is well distributed between the acts of each play, except for the *Vergonzoso* where the last act, although almost double the length of the first act, contains only half the number of running words." Clearly this statement is contradictory and cannot be corrected to correspond with the figures above by a simple adjustment.

Although Stratil and Oakley "set out to achieve a quantitative mathematical analysis", they went no further than straightforward comparisons of counts and averages, and inference by visual inspection of graphs. It is possible, however, to

attempt a statistical treatment of the frequencies of the 44 words. Including the totals for the numbers of tokens, each Act of the three plays can be compared with every other Act by means of 2 x 45 contingency tables. The program to calculate chi-square adds the numbers of occurrences of words, which give rise to expected values less than five, to the total for the remaining number of tokens in the text. The values of chi-square range from 176 for 32 degrees of freedom (d.f.) to 422 for 36 d.f. for all comparisons not involving Act III of *Vergonzoso*. (Comparisons of that Act with the eight other Acts produce values of chi-square which exceed 2000, thus confirming that the data for it is incorrect.) The values obtained, when one Act of a play is compared with another of the same play, vary from 176 to 352, while comparisons of Acts from different plays vary from 187 to 422. Without corresponding data for similar plays by a variety of contemporaneous authors, no interpretation of these results for the authorship of *Condenado* or *Burlador* is possible.

At this point Stratil and Oakley turned to what they termed a "word content analysis" of the three plays. Taking *Vergonzoso* as the basis, they listed all words at the extremes of a coefficient of difference, which unfortunately they do not define. Next, they eliminated words they believed to reflect content rather than authorship. In a curious arrangement their Table 5 followed by Table 4 combine to provide the final list of words together with their totals for each Act. The authors decided, presumably by inspection, that the words *ni* and *asi* (which occur in *Vergonzoso* more frequently than in the other two plays) reveal "the most significant difference." In the case of the former they admitted that "this difference can not easily be explained" while for the latter they claimed that "such a disparity must hold some significance." Observing its immediate context they found that, "On many occasions *asi* instigates a question and also ends an argument ... This never occurs in the other two plays, and may imply a certain immaturity in the language ... [and] may betray the hand of an author whose touch is as yet unsure." Could this be how Stratil and Oakley "established a new way of dating plays by Tirso de Molina"? If not, what is the new method they promised? – The shorter sentence length, they noted earlier, in *Condenado*?

The paper ends with eight methods of cluster analysis applied to eight Spanish plays, which include the disputed plays and *El Esclavo del Demonio* by Mira de Amescua. (The others are undefined.) Stratil and Oakley found that "the most interesting facet of the clustering is the early forming of the cluster of objects 3 and 4", where object 3 is the play above and object 4 by implication is *Condenado*. Perhaps unaware that the natural variability of authors frequently leads to such effects, they suggested that this result opens up a new area of research.

The lesson to be learnt from the experience of Stratil and Oakley is that it is imperative to plan an investigation in detail : if the experiments are not properly

designed in advance and therefore the range of essential texts is undecided when work starts, success is virtually precluded. Despite a sprinkling of comments of disarming honesty and careful wording to avoid implications of certainty, an assessment of their contribution regrettably cannot be other than harsh : Stratil and Oakley have provided only an example of how to expend much effort while unfortunately achieving nothing to further the attribution of disputed works.

Middle English Applications

In contrast with the previous studies, two related papers merit serious discussion. The first, by McColly and Weier, addressed the problem of determining whether five middle English poems are by the same author or not[10]. Their approach was to compare the texts in pairs by means of function words and certain common adjectives, adverbs and pronouns. Statistical inference depended on a likelihood ratio which was then converted to chi-square. Because the underlying theory is secure, an assessment of the method lies in evaluating the extent to which the selected features of the text conform with the prerequisites of the statistical approach and to the degree their variation reflects authorship alone.

Comparing one half of each poem with the other, McColly and Weier found all values of chi-square to be not significant at the 5 % level. In contrast, when each complete poem was compared with every other all values were highly significant. A simplistic interpretation would therefore be to accept that each poem had been written in its entirety by a separate author. McColly and Weier, however, realised that an essential stage in the derivation of their method is to validate it on works of known authorship. They therefore tested their procedure on six of the *Federalist* papers. Whether or not each of the pair of papers under test is by the same author, chi-square was found to be significant in all but one case. Furthermore, with only one exception, the values of chi-square returned, when each of a pair of papers is by a different author, were even larger. McColly and Weier therefore recognised that their results were not sufficient to formulate a case for five different poets. Nevertheless, they did make a claim for which much more evidence would be necessary. On the basis of the values of chi-square not being significant when one half of a poem is tested against the other half, they stated, "A safe conclusion, then, is that when a likelihood-ratio test reveals equal stylistic parameters ... in two works, common authorship can be safely inferred ..." Instead, in view of the usual degree of authorial variability, works by different poets could be expected at times to exhibit such similarity.

[10] William McCOLLY and Dennis WEIER : Literary Attribution and Likelihoodratio Tests : The Case of the Middle English *Pearl*-poems. *Computers and the Humanities* 17(1983), 65-75.

Realising that relative magnitudes, rather than statistical significance, may be an indicator, McColly and Weier found that three of the poems form a group while the other two are outliers. Even this did not lead to a clear conclusion because, on examination of the form of verse, the group was found to have much in common while differing from the other two poems. McColly and Weier were therefore led to suggest that "function-word rates are somewhat form-specific." Less diligent investigators, when calculating their enormous odds, fail to realise that the counts of features in texts are affected by a variety of influences on the writer. Authorship itself is but one and not always the major one.

In their introduction, McColly and Weier quoted the writers of a review as stating that a case in which only one candidate is available, is insoluable by quantitative analysis. The question as to whether the *Pearl* poems are by a single author or by many hands is equivalent by virtue of the fact that texts for comparison are not extant. The critical examination by McColly and Weier of possible explanations for their findings can be seen to lead inexorably to the same conclusion, which is that statistics cannot resolve this type of problem.

Adopting the same technique, McColly later investigated whether one of the poems, *Cleanness*, displays the lack of unity perceived by some critics[11]. Dividing the poem first where some had detected a change of authorship and then into approximately two halves, each consisting of about the same number of words, respectively, from both the previous parts, chi-square was calculated. The values obtained were 203.54 and 55.46, respectively, for 59 d.f. Studying the text, McColly decided that such a disparity was due to modality (i.e. relatively more ratiocination in one part, more narrative in the other), rather than a difference of authorship.

Both these papers are important to studies of authorship by illustrating how to avoid premature conclusions. They illustrate the sort of factors which can affect the outcome and which can sometimes combine to produce outlandish figures for odds against error. Paradoxically, the more extreme the figures, the more likely the possibility of error : abnormally large figures may indicate that the investigator has not looked for other possible causes and whether such causes could predominate. If so, they could point to a difference of authorship where none exists.

Although both these studies are much more accomplished than the two previously discussed in this article, they do contain instances of lack of care over detail. One wonders, for instance, why the total occurrences of so many words in the poem are different between the two papers when the base text is identical. Does McColly's

[11] William MCCOLLY : Style and Structure in the Middle English Poem *Cleanness*. *Computers and the Humanities* 21(1987), 169-176.

SNOBOL4 program perform accurately? For the second paper McColly has introduced the criterion of a minimum rate of one per thousand to select words. Accordingly, there is a change of nine words among the original 60. In the earlier paper, when one half of *Cleanness* was compared with the other, a value of chi-square of 55.46 for 60 d.f. was calculated. This is identical to the figure for the corresponding comparison in the later paper (based on the revised word list) and is now stated to be for 59 d.f. The division into roughly two equal parts is described as random in both papers but, in the later paper, as containing approximately equal amounts of the portions suspected as being by different authors. While it is reasonable for McColly to declare that the point of the possible change of authorship in *Cleanness* is uncertain and that any variation in its position is not enough to influence the analysis, it is unacceptable not to describe the technique used to obtain the random halves, not to explain why the counts are different and to preclude the reader from checking by omitting the figures for the calculation of chi-square as 55.46 in both papers.

The Eigenanalysis of Function Words

The method used by J.F. Burrows to distinguish between Henry and Sarah Fielding consists of selecting the 49 most frequent words in texts which are comparable with the disputed pieces and for which there is no doubt as to whether Henry or Sarah was the author[12]. From this list, personal pronouns and inflected auxiliary verbs have been excluded. A correlation matrix is calculated from occurrences in the undisputed texts. The eigenvalues and eigenvectors of this matrix are then determined. The eigenvectors associated with the three largest eigenvalues are denoted by vector A, vector B and vector C, respectively. Scaling, and taking vector B as the independent variable, vectors A and C are plotted in turn against vector B. With respect to vector B, there is no overlap between the space occupied by the ten texts written by Henry Fielding and that occupied by the ten written by his sister. Further graphs are drawn as one or more disputed texts in turn are included. Each then falls into one region or the other thereby suggesting its authorship.

Although acknowledging technical assistance, Burrows gives neither the origin of, nor the theory underlying, his method. It is, essentially, a statistical technique called *principal component analysis*. To investigate the validity of the approach, let n texts be of equal length (e.g. each consisting of 5000 words) and be denoted by $x_1, \ldots, x_n$. These n variates form a vector x. Each variate x_i consists of occurrences

[12] J.F. BURROWS and A.J. HASSALL. Anna Boleyn and the Authenticity of Fielding's Feminine Narratives. *Eighteenth Century Studies* 21 (1988), 427-53. Burrows's previous applications of his method are given in a footnote on p.430.

of the prescribed words. From these data a covariance matrix Σ is formed[13]. Its definition is denoted by

$$E(\mathbf{x}\mathbf{x}') = \Sigma. \tag{1}$$

Assume that the eigenvalues of Σ are distinct and, arranged in decreasing order of magnitude, are indicated by $\lambda_1, \ldots, \lambda_n$. The eigenvalues are found by solving

$$|\Sigma - \lambda \mathbf{I}| = 0.$$

Corresponding to the r^{th} root λ_r $(r = 1, \ldots, n)$ there is a column vector γ_r which satisfies

$$\Sigma\gamma_r = \lambda_r\gamma.$$

Let Δ be a matrix whose diagonal elements are $\lambda_1, \ldots, \lambda_n$ and whose elements elsewhere are zero. Then

$$\Sigma\Gamma = \Gamma\Delta,$$

where γ_r are the columns of Γ. Because it can be shown that Γ is an orthogonal matrix (i.e., $\Gamma' = \Gamma^{-1}$),

$$\Sigma = \Gamma\Delta\Gamma'. \tag{2}$$

The columns of Γ are the eigenvectors corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_n$.

For principal component analysis the variates $x_1, \ldots, x_n$ are transformed linearly and orthogonally into an equal number of new variates $y_1, \ldots, y_n$ which are not correlated. They are defined by

$$\mathbf{y} = \Gamma'\mathbf{x}. \tag{3}$$

The covariance matrix of y is given by

[13] The development merges general theory, based on that given by D.N. Lawley and A.E. Maxwell in *Factor Analysis as a Statistical Method* (2nd edition, 1971, Butterworth, London), with the procedure used by Burrows.

$$\begin{aligned} E(\mathbf{yy}') &= E(\Gamma'\mathbf{x}\mathbf{x}'\Gamma) && \text{(using(3))} \\ &= \Gamma'\Sigma\Gamma && \text{(using(1))} \\ &= \Gamma'(\Gamma\Delta\Gamma')\Gamma && \text{(using(2))} \\ &= \Delta. && \text{(because } \Gamma\Gamma' = \mathbf{I}) \end{aligned}$$

Because the off-diagonal elements of Δ are zero, the variates y_i are uncorrelated and the variance of y_i is therefore λ_i.

If $y_1 = \mathbf{a}'\mathbf{x}$ and $\mathbf{a}'\mathbf{a} = 1$ then y_1 would be a normalised linear combination of $x_1, \ldots, x_n$. For all y_i, premultiplying (3) by $\mathbf{c}'$

$$\mathbf{c}'\mathbf{y} = \mathbf{c}'\Gamma'\mathbf{x},$$

where $\mathbf{c}$ is defined by $\mathbf{c} = \Gamma'\mathbf{a}$.

It follows that

$$\mathbf{c}'\mathbf{c} = \mathbf{a}'\Gamma\Gamma'\mathbf{a} = \mathbf{a}'\mathbf{a} = 1.$$

Moreover,

$$\mathbf{c}'\mathbf{y} = \mathbf{a}'\Gamma\Gamma'\mathbf{x} = \mathbf{a}'\mathbf{x}.$$

Therefore, instead of the normalised linear combination of $x_1, \ldots, x_n$ denoted by $\mathbf{a}'\mathbf{x}$, it is appropriate to consider such a combination of $y_1, \ldots, y_n$, i.e. $\mathbf{c}'\mathbf{y}$.

The variance σ^2 of $\mathbf{c}'\mathbf{y}$ (i.e., $\sum_i c_i y_i$) is given by

$$\sigma^2 = \sum_{i=1}^{n} c_i^2 \lambda_i.$$

Because $\sum_i c_i^2 = 1$, σ^2 may be written as

$$\begin{aligned} \sigma^2 &= \sum_{i=1}^{n} c_i^2\lambda_i - \lambda_1 \sum c_i^2 + \lambda_1 - c_1^2\lambda_1 + c_1^2\lambda_1 \\ &= \lambda_1 + \sum_{i=1}^{n} c_i^2(\lambda_i - \lambda_1) \end{aligned}$$

Because λ_1 is the largest eigenvalue, the variance is maximized when $c_2 = c_3 = \ldots = c_n = 0$. Thus the maximum variance of y_1 occurs when $c_1 = 1$. Next, y_2 is

the normalised combination of the x-variates which has maximum variance subject to being uncorrelated with y_1. For such a combination the form is $c_2y_2 + \ldots + c_ny_n$ and for maximum variance $c_2 = 1$ with $c_3 = c_4 = \ldots = c_n = 0$. Proceeding similarly, the variance of each remaining y_i is maximized subject to its being uncorrelated with all its predecessors. The coefficients of each linear combination of $x_1, \ldots, x_n$ thus form a column of Γ, where each column is an eigenvector of Σ. Because the sum of the diagonal elements of Σ is the same as that of Δ, that is

$$\mathrm{tr}(\Sigma) = \mathrm{tr}(\Delta) = \sum_i \lambda_i,$$

the total variance of the y-variates is the same as that of the x-variates.

Frequently, principal components z_i are standardized to have unit variances i.e., $z_i = y_i/\sqrt{\lambda_i}$. Thus

$$\begin{aligned} \mathbf{z} &= \Delta^{-1/2}\mathbf{y} \\ &= \Delta^{-1/2}\Gamma'\mathbf{x}. \end{aligned}$$

In authorship studies, what the components represent is not clearly defined. In Burrows's work the largest eigenvalue accounts for 90 % or more of the sum of the eigenvalues. This predominant component could be interpreted as a 'semantic factor', in the sense that words of the type prescribed always occur with comparable patterns of frequencies : for instance, *the* always occurs much more often than *at* because, independent of authorship or any other factor, the ratio of their use is constrained so that the text is meaningful. This is reflected in the elements of the corresponding eigenvector, which do not form a discernable pattern. Indeed, all the texts (i.e., the x-variates) contribute similarly to this component.

In contrast, the elements of the eigenvector corresponding to the next largest eigenvalue are of one sign for Henry Fielding's work and the opposite sign for Sarah's. Because the texts are of the same genre and originate in the same culture, the associated component would appear to be authorship. No obvious pattern emerges in the elements of the remaining vectors so they do not seem to possess readily identifiable meanings.

Because only the relative values of the elements of the eigenvectors are important in Burrows's procedure, any suitable scaling factor may be introduced to replace $\Delta^{1/2}$. There are, however, other difficulties. According to Lawley and Maxwell, principal component analysis is most useful when the variates are all measured in the same units. When the texts are of different lengths, scaling factors are implicitly introduced so this application is of a type where they find its use "more difficult to justify". Although Lawley and Maxwell do not give their endorsement,

to overcome this difficulty they point out that in educational and psychological work it is common practice to standardize each variate and then to replace the covariance matrix by a correlation matrix. This is the approach which Burrows adopted. Its implications are examined next.

Replacing the values for each variate by

$$z_i = \frac{x_i - \overline{x}_i}{\sigma_i},$$

where $\overline{x}_i$ and σ_i are their mean and standard deviation, respectively, creates new variates of zero mean and a standard deviation of unity. It can be shown that if N is the number of values in each variate, the correlation matrix R, calculated from

$$\rho_{ij} = \frac{N\Sigma x_i x_j - \Sigma x_i \Sigma x_j}{[\{N\Sigma x_i^2 - (\Sigma x_i)^2\}\{N\Sigma x_j^2 - (\Sigma x_j)^2\}]^{\frac{1}{2}}}$$

is unaffected by such a change.

Because the elements of the covariance matrix are determined from

$$\sigma_{ij}^2 = \frac{1}{N}\sum(x_i - \overline{x}_i)(x_j - \overline{x}_j),$$

writing σ_{ii}^2 as σ_i^2 and σ_{jj}^2 as σ_j^2, the elements ρ_{ij} of R may be expressed in terms of covariance and standard deviations :

$$\rho_{ij} = \frac{\sigma_{ij}^2}{\sigma_i \sigma_j}$$

Moreover, the elements of the covariance matrix formed from the corresponding normalized variates, cov_{ij}, can be shown to be given also by the same expression :

$$\mathrm{cov}_{ij} = \frac{\sigma_{ij}^2}{\sigma_i \sigma_j}$$

It follows that use of a correlation matrix is equivalent to substituting normalised variates for the original variates and proceeding in accordance with the accepted theory as outlined above. Nevertheless the ratio of corresponding elements is altered by this step, so the effect on the outcome is not easily predictable and is liable to be more severe in some cases than in others.

Despite this deviation from the underlying theory, Burrows's results are impressive. To test the method I applied it to the Acts of plays by Middleton and Tourneur to identify which is more likely to have written *The Revenger's Tragedy* printed anonymously in 1607. Because the Acts are of differing lengths a correlation matrix

was used. Again the eigenvector corresponding to the second largest eigenvalue appeared to contain information on authorship. The area occupied by Middleton, however, was found to extend into that of Tourneur. Moreover, Acts II and V of *The Revenger's Tragedy* also fell into Tourneur territory, while the remaining Acts were located in the region of Middleton which was closest to Tourneur's.

In contrast, when all the plays were divided into blocks of 5000 consecutive words and a covariance matrix was calculated, such ambiguity was eliminated : all blocks of Middleton were separated from those of Tourneur. Moreover, no part of *The Revenger's Tragedy* extended into the region occupied by Tourneur's work. And, in a further comparison, Middleton and Marston were distinguished equally unambiguously with *The Revenger's Tragedy* falling clear of the latter's territory[14]. Applied with due attention to the underlying theory, the outcome of the method is therefore consistent with scholarly study and perception.

A further advantage of applying principal component analysis to the resolution of problems of authorship is that no hypothesis need be made about x_i.

Principles of Attribution

Writing in a forensic context RW Bailey considers that there are (at least) three rules to define the circumstances necessary for attribution[15] :

1. the number of putative authors should constitute a well-defined set;
2. the lengths of the writings should be sufficient to reflect the linguistic habits of the author of the disputed text and also those of each of the candidates;
3. the texts used for comparison should be commensurable with the disputed writing.

In scholarly work, when a possible author cannot be included (if, for example, suitable specimens of his writing are not extant), a curtailed investigation is still valid. The result can then be declared either in a relative form, such as the disputed work shows a greater affinity with one author than the others or that none of the group tested provides a satisfactory match. When a number of independent stylometric approaches leads to the same conclusion, any subsequent literary or historical

[14] To be published.

[15] Richard W. Bailey : Authorship Attribution in a Forensic Setting. In *Advances in Computer-Aided Literary and Linguistic Research. Proceedings of the Fifth International Symposium on Computers in Literary and Linguistic Research* (Ed. D.E. Ager, F.E. Knowles and Joan Smith), Birmingham, 1979, pp. 1–15.

investigation can be simplified by concentrating on the possibilities (including multiple authorship) still unresolved.

Ellegård's method appears to be particularly attractive because it seems to overcome Bailey's first rule[16] : it promises the capability to demonstrate that a text is by a specific writer independent of how many potential authors were available at the time. Lake adapted this approach for his study to ascertain the scope of the Middleton canon[17]. Unfortunately, when the details were investigated it was found that the premises of the method were not necessarily fulfilled and that this was the likely origin of the very small fractions which he had calculated[18].

A statistical study of disputed authorship consists essentially of :

1. A definition of the problem and a list of possible authors.
2. The selection of appropriate texts with which to compare the disputed text.
3. The choice of classes of tests and the statistical methods to be applied.
4. If appropriate to the nature of the test, the determination of the individual features which distinguish each of the authors when taken in pairs.
5. Validation of the procedures for the particular problem at hand.
6. Comparisons of the disputed text with works by each of the possible authors (taken in pairs, if appropriate).
7. Analysis of the results in the light of all information available on the particular problem.

Authors are taken in pairs when it would be unusual to find one or more features which reliably distinguish any writer from all his contemporaries. Moreover, attribution studies should not be performed in isolation: one item of external evidence can overturn all such internal evidence. More generally, authorship studies should observe general principles, the most important of which are[19] :

1. The onus of proof lies entirely with the person making the ascription.
2. The argument for adding something to an author's canon has to be vastly more stringent than for keeping it there.
3. If doubt persists, an anonymous work must remain anonymous.

---

[16] A. Ellegård : *A Statistical Method for Determining Authorship* (Gothenberg Studies in English, Vol. 13). The University of Gothenberg, 1962.

[17] See note 3, above.

[18] See note 5(iv), above.

[19] M.W.A. SMITH : *Edmund Ironside* and Principles of Authorship Attribution. *The Shakespeare Newsletter* 38(1988), p.50.

4. Avoidance of a false attribution is far more important than failing to recognise a correct one.
5. Only works of known authorship are suitable as a basis for attributing a disputed work.
6. There are no short-cuts in attribution studies.

Conclusions

Turning to the investigations critically examined in this paper, Baker's suggestion of an author's *pace* must be set aside because he failed to validate it sufficiently rigorously. Stratil and Oakley's steps towards the attribution of Spanish plays are deficient in each of the seven components above of statistical studies. Accordingly they paid the price in failing to make any progress whatever. On the other hand McColly and Weier considered the nature of their problem carefully but because none of the seven components of the study can be performed satisfactorily, it was impossible for them to reach a conclusion. Their technique, however, is based firmly in statistical theory but its main weakness appears to be its reliance on the assumption that the words on which it is based are Poisson distributed. This condition would often seem to be violated[20] and thus may reduce the suitability of the method for general use. Further study, however, is needed to ascertain if such deviations from Poisson distributions merely degrade the outcome somewhat, or if they are likely to invalidate the result.

By far the most promising of the methods examined is the use of function words in conjunction with the technique of principal component analysis, provided that a covariance matrix is used rather than its convenient replacement by a correlation matrix. Consequently, the data must be obtained from blocks of text of equal length.

This survey of some recent proposals for, and examples of, attribution by statistics has illustrated that stylometry can be a serious science when investigators proceed carefully and do not attempt too much. In contrast, if quick and easy answers are sought, the outcome is likely to be of little value and worse, may induce literary scholars, unequipped to assess the details, to reject all such studies. Nevertheless, the approach adopted in this paper reveals that the alert non-scientist, armed only with common-sense, should frequently be able to detect an unreliable attribution.

---

[20] Fred J. DAMERAU : The Use of Function Word Frequencies as Indicators of Style. *Computers and the Humanities* 9(1975), 271-80.

Acknowledgement

I am grateful to my colleague Dr N.T. Neill for useful discussions on principal component analysis and for making available his flexible program to determine eigenvalues and eigenvectors.