Forensic Stylometry: A Review of the Cusum Controversy

David I. HOLMES and Fiona J. TWEEDIE

Résumé. Les graphes de sommes cumulatives des différentes classes de mots ont récemment été utilisés pour mettre en doute, lors de procès, les aveux de l'accusé. Cet article se propose d'examiner d'une part, la validité de cette technique, et d'autre part, l'hypothèse sur laquelle cette technique est développée. Cet article examinera les travaux de recherche relatifs à l'application des graphes de sommes cumulatives. L'efficacité d'une autre procédure reposant sur la pondération des sommes cumulatives est également explorée. Notre conclusion remet en cause la crédibilité des indicateurs de l'identité de l'auteur.

Keywords: Cumulative sum, stylistics, forensic linguistics, word usage, weighted cumulative sums.

Mots-clés : Somme cumulative, stylistique, expertise linguistique légale, utilisation de mots, poids de sommes cumulatives.

1. Introduction

Cumulative sum charts or "cusum" charts are primarily used in industrial processes and quality control monitoring, but have also found application in authorship attribution studies. Bee (1971); Bee (1972) first proposed using cusum charts of verb frequencies as a test of authorship for biblical texts. Michaelson *et al.* (1978) employed cusum charts of sentence lengths to characterize Greek writers and, more recently, Andrew Morton has proposed an authorship test which uses cusum charts of various word classes. Three

✉ David I. HOLMES; Department of Mathematical Sciences; University of the West of England, Bristol; Coldharbour Lane, Frenchay; Bristol BS16 1QY (United Kingdom).
Fax: +44 117 976 3860 E-mail: di-holme@csm.uwe.ac.uk

✉ Fiona J. TWEEDIE; Department of Mathematical Sciences; University of the West of England, Bristol; Coldharbour Lane, Frenchay; Bristol BS16 1QY (United Kingdom).
Fax: +44 117 976 3860 E-mail: fj-tweed@csm.uwe.ac.uk

internally published reports [Morton and Michaelson (1990); Farrington and Morton (1990); Morton (1991)] provide detailed accounts of the approach to interpreting cusum charts in the context of authorship attribution.

Since 1990 Morton's cusum technique has been used by defence lawyers seeking to cast doubt over confessional statements allegedly made by their clients. Examples include the appeal in London of Tommy McCrossen in July 1991, the trial in November 1991 of Frank Beck in Leicester, the trial of Vincent Connell in Dublin in December 1991 and the pardon for Nicky Kelly from the Irish government in April 1992. In the McCrossen case, cusum evidence which cast doubt over the authenticity of the defendant's confession played a vital role in persuading the appeal judges that he had been wrongly convicted. In most of the cases in which cusum charts have been used in the courts, they have been said to reveal that material had been inserted into the statement that was not in the verbal style of the appellant and that the findings supported the case that the confessions were modified by police officers to make them more clearly incriminating of the defendant. Cusum charts were also used in the enquiry concerning academic standards in the Department of Philosophy at the University of Wales, Swansea, [Farringdon (1992)].

The cusum technique has also been used by other people in their investigations. Taves (1991) uses cusums to analyse the authorship of the *Book of Mormon*. His initial graphs seem to discredit the technique, as texts from purportedly different authors produced apparently consistent cusum graphs. When comparing text from the biblical *Genesis* against text from *The Book of Abraham*, a Mormon scripture, he notes that

> "all texts do not reveal multiple authorship even when we know this is the case",

yet Taves claims to be guided by Morton in his stylometric investigations. This directly contradicts the 100 % accuracy claims of Farringdon and Morton. A more detailed study of the authorship of Mormon scripture can be found in Holmes (1992). Davies (1990) looks at a pamphlet attributed to Defoe and, after claiming to validate the technique on similar authors, concludes that Defoe was not the author of the disputed work.

Serious doubts are now being raised about the validity of Morton's cusum as a forensic technique and these doubts have been aired both on British television (BBC's *Tomorrow's World* and Channel 4's *Street-legal*) and in the press [Campbell (1992), Matthews (1993a)]. Popular journals such as *New Scientist* have also added publicity to what has become known as the "cusum controversy" [Matthews (1993b)].

Whereas proponents of the cusum test put forth "proof by example", offering up dozens of examples in which they claim the method demonstrates evidence of multiple authorship, recently several papers have been published criticizing the technique [Canter (1992); Hilton and Holmes (1993); Hardcastle (1993); de Haan and Schils (1993b); Sanford *et al.* (1994) and, in addition, a report has been sent to the British Crown Prosecution Service]. Critics of cusum are concerned over the ease with which it has been accepted by courts and the deeper question about how juries and the judiciary can be expected to assess scientific evidence. In the United States, the judicial system has dealt with new scientific techniques by subjecting them to the "Frye Test"; this permits only testimony based on "generally accepted" scientific techniques to be put before a jury.

In this paper we present a description of Morton's cusum technique and look at the main criticisms directed against it. We then review the current literature proposing alternative methods of analysis.

2. The Cusum Technique

Of central importance to the preparation and interpretation of cusum charts in this linguistic context is Morton's claim that each person has a unique set of habits which he or she follows consistently whenever communicating, whether through the written or spoken word. These habits are quantifiable in that they are particular components of that person's sentences. Examples of such habits are usage of nouns, short words (*i.e.* words of 2 or 3 letters) and words beginning with a vowel, known as initial-vowel words. Further classes of habits may be formed by combining two or more of these simple classes; Morton often uses the class of (short words + initial-vowel words). It is proposed that the rate of occurrence of a habit is so consistent for each individual that any distinct variation in the proportion of occurrences of the habit within a sample of sentences is *prima facie* evidence that the sentences are the utterance of more than one person. Similarly, a disputed text, containing words of a particular class occurring at a rate which is consistent but which differs significantly from specimen texts of a candidate author, would be judged to be of different authorship.

The cusum technique requires that one generates and compares two cusum charts, one for the sentence lengths and one for the number of times

the habit in question occurs in each sentence. If we have n sentences, define w_r, where $r = 1 \ldots n$, to be the number of words in sentence r, and

$$\overline{w} = \frac{1}{n} \sum_{r=1}^{n} w_r ,$$

then plot

$$c_i = \sum_{r=1}^{i} (w_r - \overline{w}) \qquad (1)$$

against i for $i = 1 \ldots n$. Define also x_r, where $r = 1 \ldots n$, to be the number of occurrences of the habit under scrutiny in sentence r, and

$$\overline{x} = \frac{1}{n} \sum_{r=1}^{n} x_r ,$$

then plot

$$h_i = \sum_{r=1}^{i} (x_r - \overline{x}) \qquad (2)$$

again against i, for $i = 1 \ldots n$.

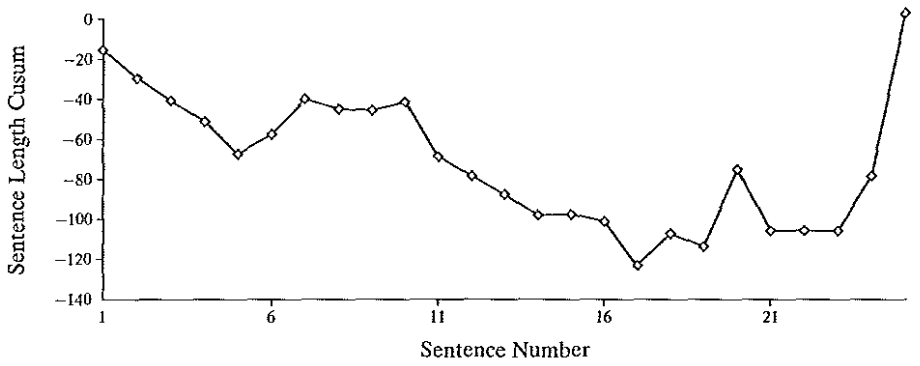

Fig. 1.– Cusum plot of sentence lengths from Austen's *Northanger Abbey*

Figure 1 shows the cumulative sum plot for the lengths of the first twenty-five sentences of *Northanger Abbey*, where the mean sentence length is 36 words and the vertical scale is the cumulative sum of deviations, in words, from this mean. We must note that it is also possible to perform cusum analysis using a metric other than sentence length. The length of any periodic structure in a text may be used; for example, the line in poetry. Morton and Michaelson (1990) give several examples of such metrics. Figure 2 plots, for the same

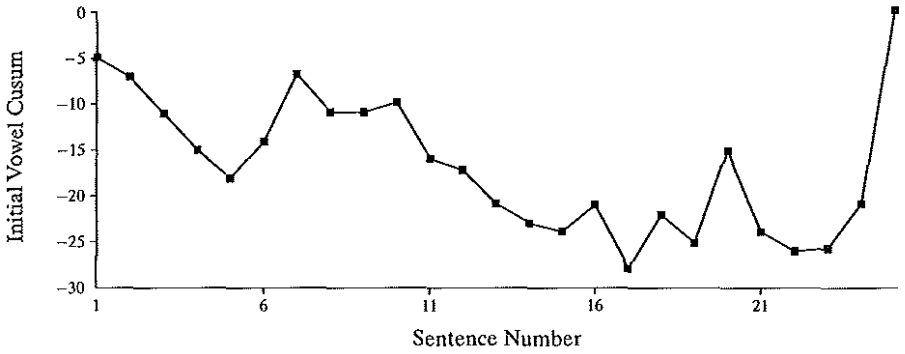

Fig. 2.– Cusum plot of initial-vowel words from Austen's *Northanger Abbey*

Northanger Abbey example, the cusum chart for words whose initial character is a vowel. We must now superimpose the two plots, and here Morton applies a scaling factor to the values in the habit cusum plot before superimposing it upon the sentence-length cusum plot. The scaling factor he uses is the ratio of the maximum to minimum cusum value ranges for the individual plots. Hardcastle (1993) contends that this scaling factor is imprecise and the way in which Morton's factor is calculated means that it is a function of the average sentence length and habit words rate only within that portion of the text represented by the cusum chart points lying between the highest peak and the lowest trough. Figure 3 shows the final cusum plot for the *Northanger Abbey* example, with the sentence length and habit cusums superimposed. They take a similar shape and the plot does appear to support Morton and Michaelson's claim that Jane Austen is consistent in her usage of initial vowel words.

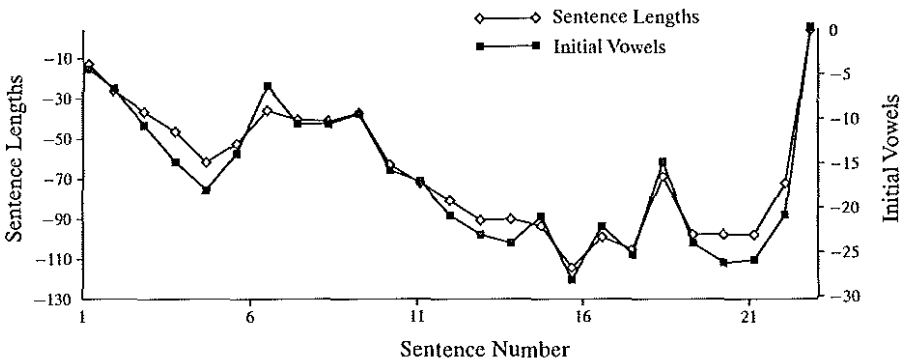

Fig. 3.– Cusum plot of sentence lengths and habit words from Austen's *Northanger Abbey*

The central premise for interpreting cusum plots such as Figure 3 is that these two values (sentence-length and the number of habit words per sentence) should parallel each other in the utterances of any one person. A significant divergence between the two values would, Morton claims, demonstrate different authorship for part of the text since it would indicate a difference in the rate of use of the habit. He makes no use of standard statistical measures to compare the two sets of data values and insists that visual inspection of the charts is sufficient to decide whether there is intruded material or not. Morton (1991) states that if the

> "variations [measured visually] within the charts are as large as the variations between them this may be taken as evidence of a consistent habit".

He advances the criterion, therefore, that when the maximum vertical separation between the sentence-length cusum line and the habit cusum line is less than the separation between any two successive points within either of the lines, the rate of use of the habit is consistent.

Hardcastle (1993) argues that this criterion makes no sense at all. The maximum separation between points within a line is determined usually by the longest sentence in the text, and, because cusum portrays cumulative deviations, inter-line separations are more likely to be greatest near the middle of the chart than near to either end. Indeed, if we assume, for the moment, that sentence lengths are independent and that

$$\mathrm{var}\left(\sum_{r=1}^{n} w_r\right) = \sigma^2, \tag{3}$$

then

$$\mathrm{var}\left(\sum_{r=1}^{i} w_r\right) = \frac{i}{n}\sigma^2, \tag{4}$$

and

$$\mathrm{var}\left\{\sum_{r=1}^{i}(w_r - \overline{w})\right\} = \frac{i}{n}\left(1 - \frac{i}{n}\right)\sigma^2. \tag{5}$$

Thus the variation is not constant along the sequence of the cusum, Equation 5 shows that it is much greater in the middle than near either end. This naturally affects the ability of the cusum chart to detect changes in authorship.

Another related defect in the technique emerges when the starting point of the chosen text is changed. If, say, there was a change in habit rate in sentences 50–70 of a 100 sentence text, and a cusum chart was plotted of sentences 20–90, then it is likely that a divergence will appear on the chart.

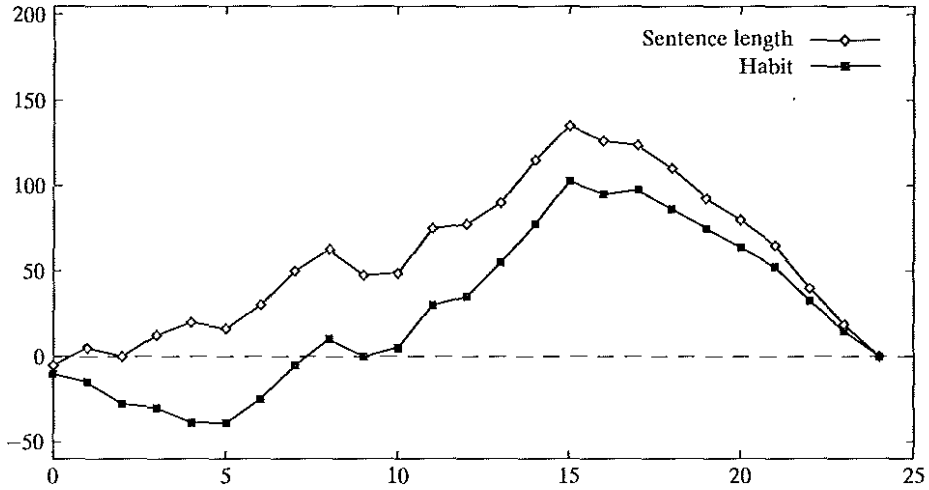

Fig. 4.– Change in habit rate between sentences 1 and 5

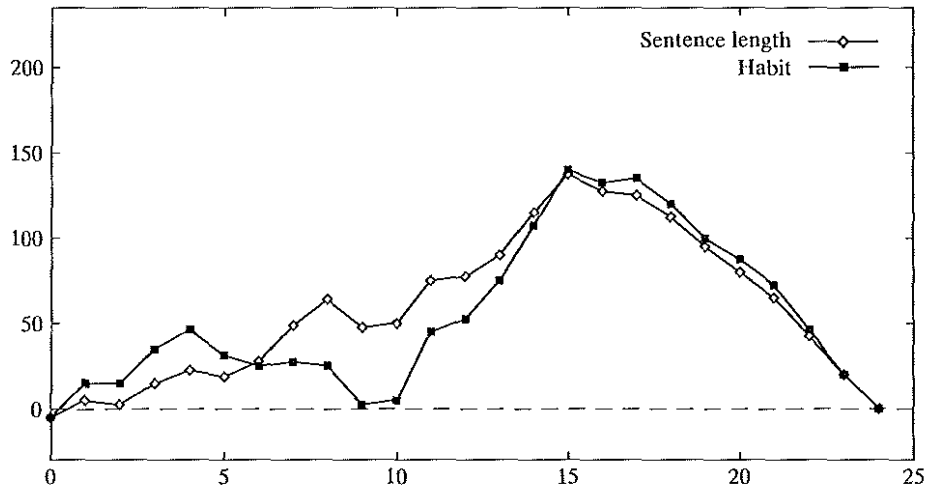

Fig. 5.– Change in habit rate between sentences 6 and 11

However, if the finishing point was moved to sentence 70 it is highly unlikely that any visual effect would be observed. Figures 4 to 8 illustrate how a change in habit rate affects the cusum plot when that change is in different places in the chart. In each chart the sentence lengths are taken from Jack London's *The Sea Wolf*, the simulated habit fixed at 0.50 of each sentence

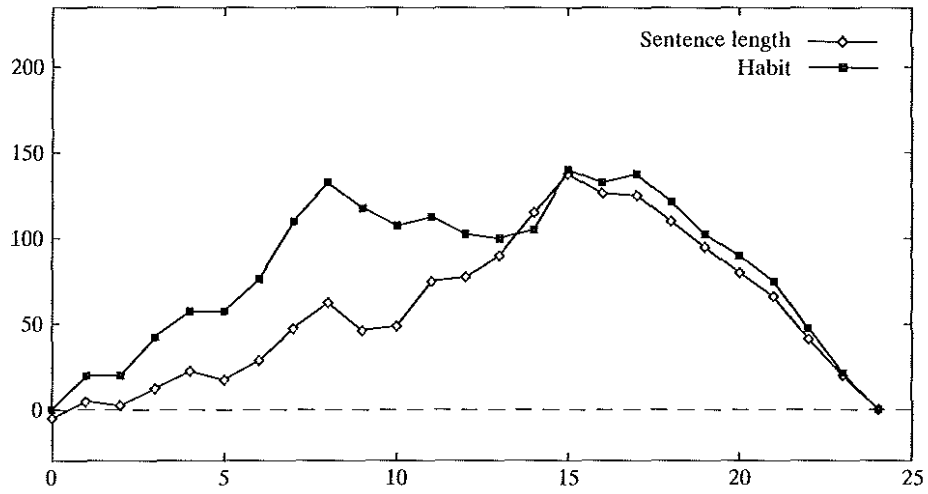

Fig. 6.– Change in habit rate between sentences 11 and 16

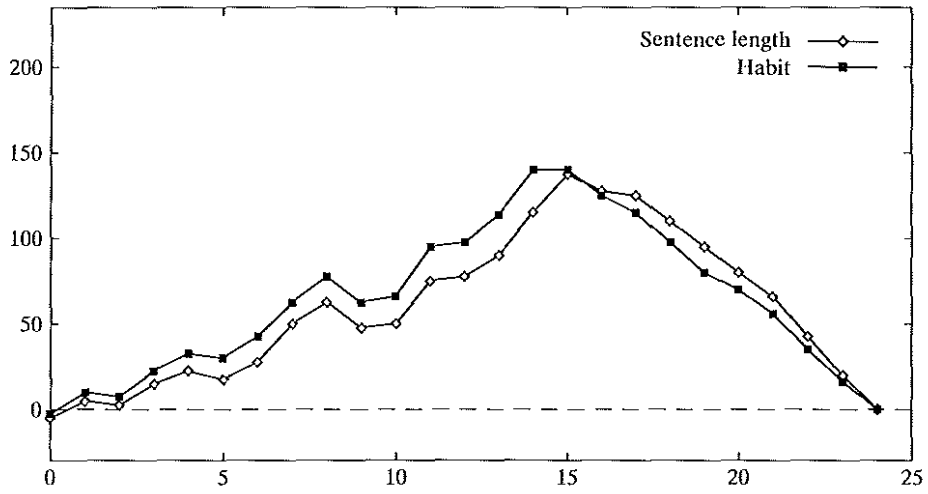

Fig. 7.– Change in habit rate between sentences 16 and 21

length, then rounded to the nearest integer. In the initial chart, Figure 4, the first five sentences have a habit rate of 0.25, simulating a change in authorship at sentence 6. It can be seen that the charts diverge over the first five sentences, before gradually converging to the end of the sample. Figure 5 shows that when sentences 6 to 10 have a habit rate of 0.25, the plots again diverge over

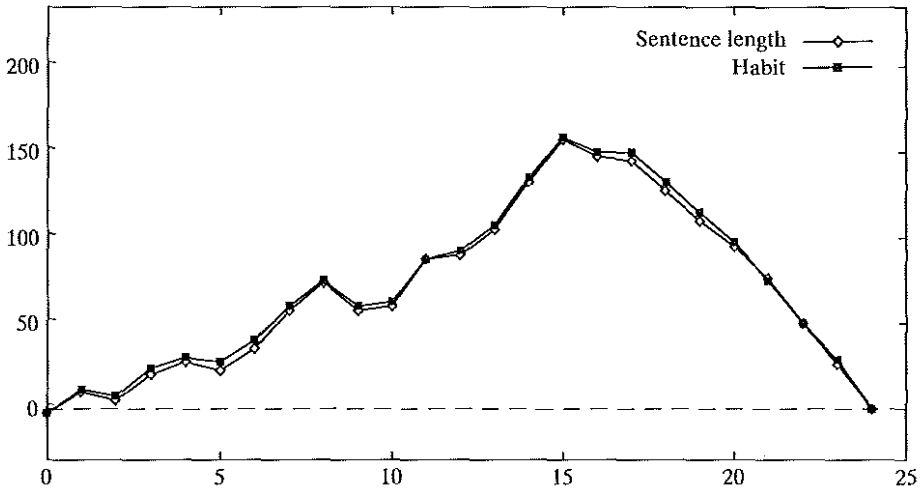

Fig. 8.– Change in habit rate between sentences 21 and 25

this section, before converging to meet at the end of the sample as they are constrained to do. Figure 6 shows the similar case when sentences 11 to 15 have the 0.25 habit rate. However, when the habit rate is changed between sentences 16 and 21, as in Figure 7, there is a very slight divergence between the two plots, while in Figure 8 the charts are almost exactly coincident, indicating that there is no change in the use of the simulated habit. Hence, a change in the starting point of the cusum chart can affect whether or not a variation is detected. A technique which can indicate changes of habit rate only in the first half of the text can hardly be considered consistent. We believe that this is an important aspect that should be investigated further. Work is continuing and we hope to publish a separate report in due course.

To decide whether the two charts "match" on a cusum plot, Morton advocates the use of a transparent overlay of the habit cusum graph so that it may be moved around with respect to the sentence-length cusum graph. Material that does not match is claimed to be by a different author. Morton claims that this approach is easy for juries to understand, yet the procedure is extremely subjective and is open to *ad hoc* interpretation. Judgment is also hampered by the use of different vertical scales in the cusum graphs he presents. An example of this may be found in the analysis of the taped telephone conversations, purported to be between the Prince of Wales and Camilla Parker-Bowles (his alleged mistress), as well as recorded telephone speech purported to be that of the Princess of Wales, [Harrison (1993)]. Here

it is claimed that the cusum graphs derived from the conversations of the Prince and Princess of Wales show "significant divergences" whilst that for Camilla Parker-Bowles does not. The vertical scale for Camilla's cusum graph is compressed to approximately half that of the other two!

Morton claims that people are consistent in their rates of use of particular classes of words, irrespective of the type of utterance, written or spoken, personal or formal, conversation or interview etc. To overcome objections to this claim, Morton and Farrington (1992) use material from the case of the "Birmingham Six" and compare part of the record of an interview of one of the accused conducted by an assistant chief constable and a personal letter written by the accused many years later. Hardcastle (1993) contends that there are many linguistic differences between writing and speech and adds

> "... the claim that the [cusum] method is insensitive to these is a remarkable one."

Sanford *et al.* (1994) investigate this further. They carried out an experiment in which students were asked to write and describe in words a short video that they had just watched. According to the *lexical density* hypothesis, see for example Halliday (1990), one would expect spoken language to have a higher density, or number, of function words such as prepositions and connectives, and a lower density of content words such as nouns.

Several habits were enumerated from the students' written and spoken language; the number of two and three letter words, the number of three and four letter words, the number of words beginning with a vowel, the number of words beginning with a vowel that were not two or three letter words and nouns in each sentence. The ratio of each of these habits to the total number of words in the text was then examined for both the written and spoken versions. Sanford *et al.* (1994) found that there were statistically significant differences between the written and spoken ratios for each habit except the number of words beginning with a vowel that do not have two or three letters. In the written versions the number of nouns is higher, while the rest of the habits occur more frequently in the spoken versions. These findings are consistent with the lexical density hypothesis rather than with Morton's claims. It would appear that Morton's examples of this are one-off cases rather than serious proof of this aspect of cusum charts.

Morton and Farrington (1992) also deal with the question of whether consistency of the habit is still manifested if samples are examined from the same writer but are separated by many years and deal with different subjects. Using material written by Morton himself between 1961 and 1980, the authors contend that when samples from one source are taken together, the rate of

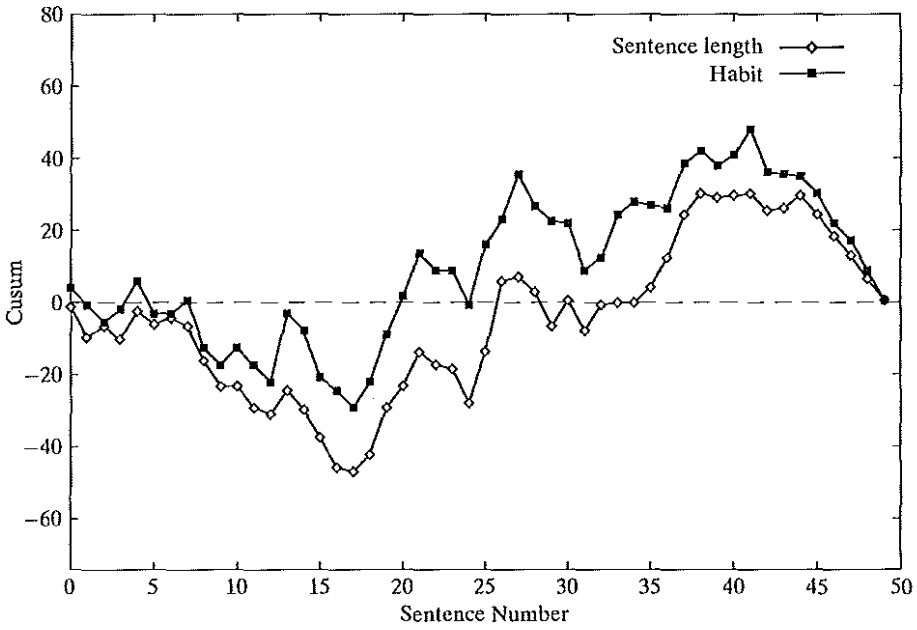

Fig. 9.– Asimov's *Foundation* and *Fourndation and Earth*

occurrence of the particular class of words under study remains unchanged. However, this consistency through time does not appear to be the case for all authors. Figure 9 shows the cusum plot of sentences from Asimov's *Foundation* saga where the habit is initial-vowel words. The first twenty-five sentences are taken from *Foundation*, written in 1960, and the next twenty-five from *Foundation and Earth*, written in 1986. As can be seen in Figure 9, there appears to be a discrepancy between sentences ten and forty-five, indicating that Asimov did not use initial-vowel words at the same rate in 1960 as he did in 1986.

One of the justifications put forward by Morton and Michaelson (1990) for the use of cumulative sums is that language is periodic. De Haan and Schils (1993a) show that sentence lengths in English are indeed correlated. This was an important motivation in the use of time series analysis by Jamieson and Aitken (1993). However in the dispute at the University of Wales, Swansea, Farringdon (1992) picks and chooses the sentences he wishes to include from a multiple choice examination paper that can not be described as a continuous utterance.

In another case, when independently analyzing the same piece of text, Farringdon and Morton choose different sentences to omit. One sentence is omitted by Farringdon because he claims that it is a list, yet Morton includes it in his analysis and other similarly structured sentences are left in the text. This gratuitous editing of texts contradicts one of Morton and Michaelson (1990)'s motivations as well as ignoring lexical aspects of continuous written text.

3. Combining sources from different samples

Proponents of the cusum test claim that it is possible to discriminate between two texts written by separate authors who have consistent, but different habits by concatenating the texts and calculating sentence-length and habit cusum charts for the combined text. After the charts are scaled and their plots superimposed, there should be a significant discrepancy between the plots near the point where the two texts were concatenated; this discrepancy may or may not continue for the rest of the plots. When one of the texts in question is much smaller than the other, Morton and Michaelson (1990) suggest inserting the smaller text into the middle of the larger text.

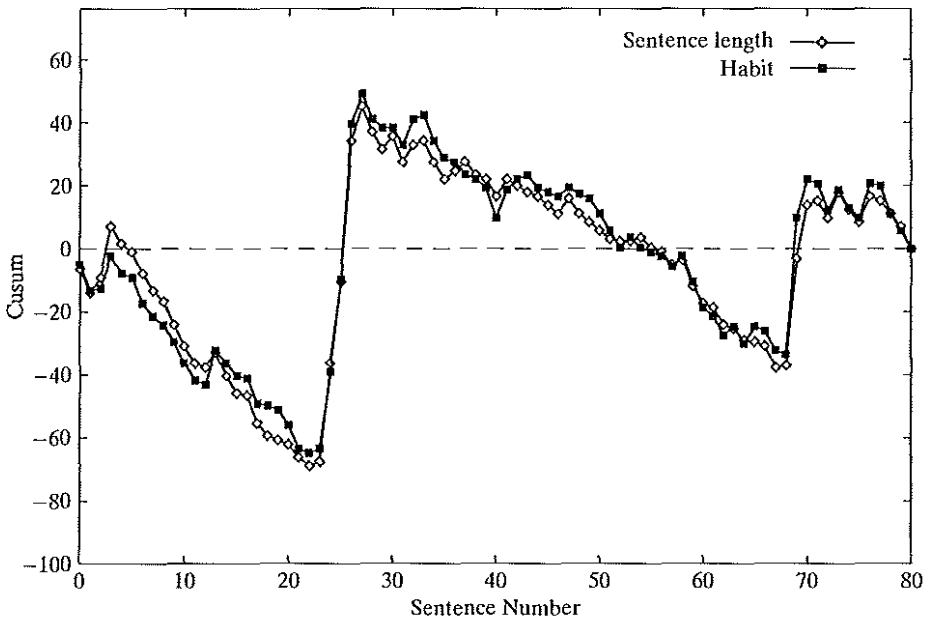

Fig. 10.– Heinlein's *Stranger in a Strange Land* and Austen's *Sense and Sensibility*

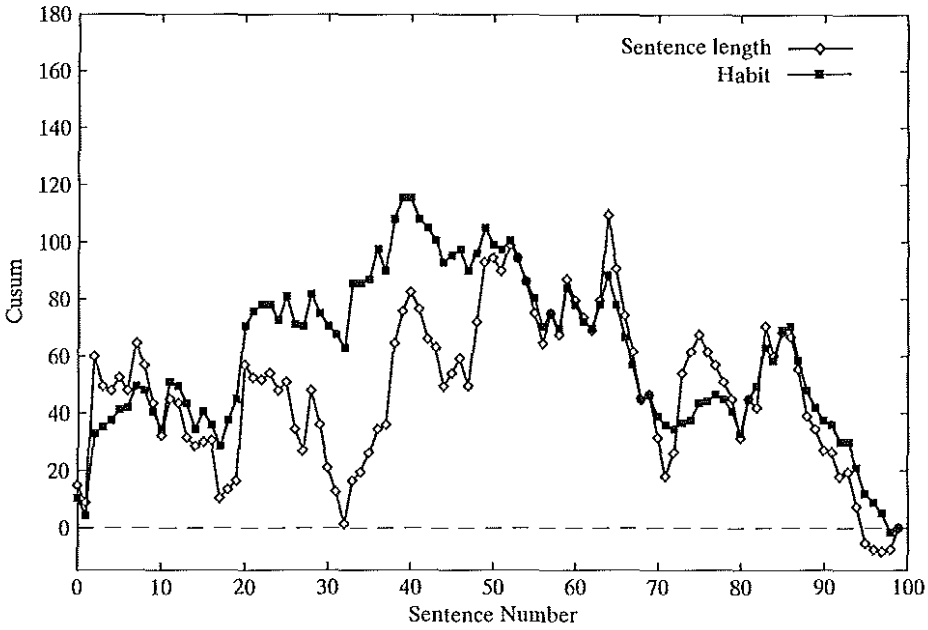

Fig. 11.– Conan Doyle's *A Case of Identity*

This procedure, however, can only demonstrate that the two texts have different average rates of use of the habit under scrutiny; it is not a reliable test of authorship. Hardcastle (1993) shows how different specimen texts from the same author can have significantly different average habit rates although, when taken individually, the texts have closely corresponding cusum sentence-length and habit lines. Figure 10 shows how texts from different authors can result in an apparently consistent cusum plot. Here the first 23 sentences are from Robert Heinlein's *Stranger in a Strange Land*, the next 5 from Jane Austen's *Sense and Sensibility* and the remainder from the same passage of Heinlein. The habit under study is the rate of occurrence of short words. It can be seen that Austen's sentences are on average much longer than Heinlein's, yet no divergence in the plot can be found near the joins. Figure 11 illustrates the converse, showing the cusum plot for the first 100 sentences of Conan Doyle's *A Case of Identity*, with initial-vowel words as the habit. There are many discrepancies in this plot, leading one to believe that it is of multiple authorship.

The subjectivity involved when visually interpreting these joint cusum charts is well illustrated by the following examples. Figure 12 shows text

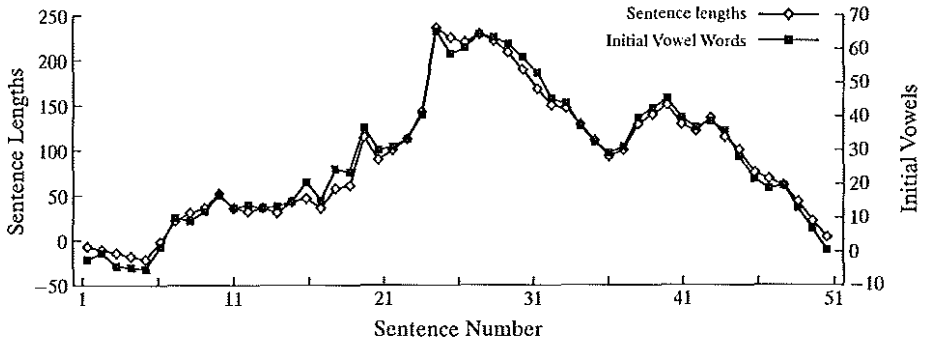

Fig. 12.– Austen's *Northanger Abbey* and *Mansfield Park*

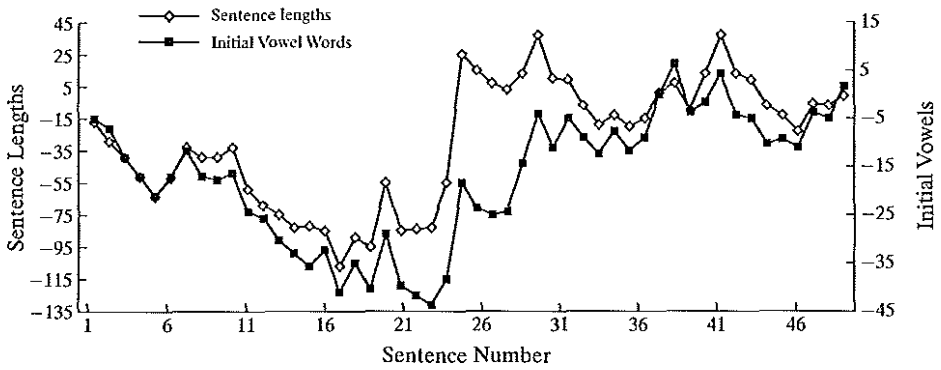

Fig. 13.– Austen's *Northanger Abbey* and Fitzgerald's *The Great Gatsby*

from *Northanger Abbey* concatenated with text from Austen's *Mansfield Park*, the habit under study being initial-vowel words. There appears to be no significant discrepancy near sentence twenty-five where the two samples join. The cusum plot in Figure 13, which compares the usage of initial-vowel words in *Northanger Abbey* and F. Scott Fitzgerald's *The Great Gatsby*, does appear to show a significant discrepancy, and using Morton's method of transparent overlays this discrepancy may be localized to lie between sentences twenty-three and twenty-eight.

Figures 14 and 15 concern the change of authorship in the novel *Sanditon*. Jane Austen died while writing *Sanditon*, prematurely ending the novel at sentence seventy-three of Chapter 11. Many years later, the novel was completed by another writer and published as *Sanditon*, by *Jane Austen and Another Lady*. Figure 14 shows the cusum chart for sentence-length and initial-vowel

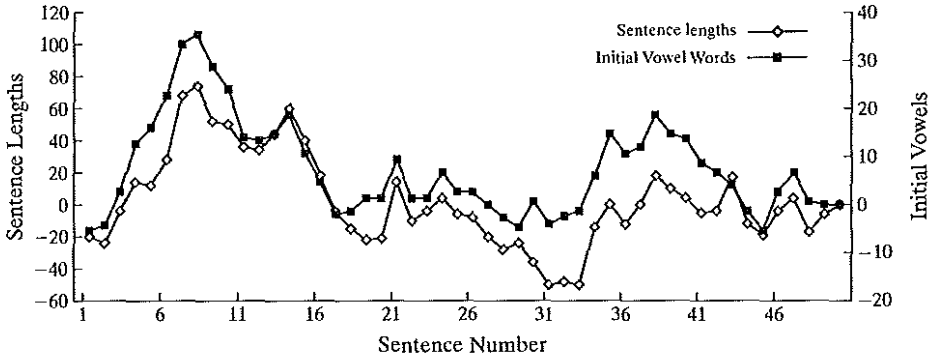

Fig. 14.– Austen's *Sanditon*

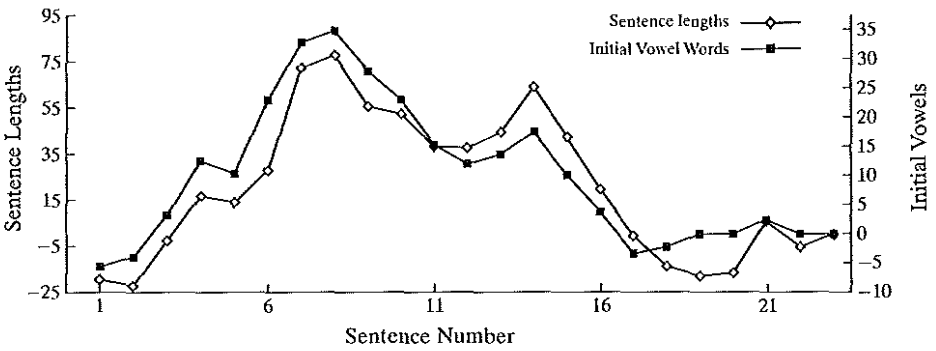

Fig. 15.– Detail of Austen's *Sanditon* from Figure 14

words for the last fifty sentences of Chapter 11; the first twenty-three were written by Austen and the final twenty-seven by the Other Lady. Morton and Michaelson (1990) claim that the discrepancy appearing at the point where authorship changes is clear and "irreconcilable". There does appear to be a discrepancy between the sentence-length and initial-vowel plots that begins around sentence twenty-three, but there are also other discrepancies any of which may also be interpreted as "significant". If Morton and Michaelson's definition of a significant discrepancy is applied to the cusum chart constructed for the twenty-three sentences written by Austen alone (shown in Figure 15), one could conclude that the discrepancy occurring between sentences eleven and twenty-one indicates a change of authorship!

It is interesting to note that the scale of the *Sanditon* cusum chart presented by Morton and Michaelson is larger than that of Figure 14. Their

scale ranges from 150 to −150 for the sentence-lengths and from 52 to −52 for the initial-vowel words, thereby making the discrepancies before sentence twenty-three look smaller and less significant. It should also be noted that Morton and Michaelson (1990) recommend

> "a general rule is that the vertical height of the chart should be about 30 to 40 % of the width of the chart."

However, in their cusum plot of *Sanditon*, the vertical height is over 43 % of the width while in Figure 14, the vertical height is about 33 % of the width of the cusum plot. It could be concluded that Morton and Michaelson's interpretation of the plot was biased, reflecting their knowledge about the text in question. This example typifies Morton's somewhat cavalier approach to scaling, well illustrated by the comment [Morton and Farringdon (1992)]:

> "The experimental approach is also the safe guide to fixing the scales of the charts."

This procedure of combining two texts into a single cusum chart is also used by Morton in a forensic context [see Morton and Farrington (1992)]. Here, confessional statements are concatenated with sentences uttered by the officer recording the interview and also with genuine utterances of the accused. In the McCrossen case, Morton and Farringdon contend that when sentences from the confessional statement were joined with a similar number of sentences from the accused, the charts "unmistakably separated". Clearly, the definition of what constitutes a significant discrepancy is central to the interpretation of cusum plots.

4. Review of Alternative Methods

In this section we aim to review alternative methods of analysing cusum charts that have been published recently. We examine work by Canter (1992) who uses Spearman's rank correlation coefficient; Hardcastle (1993), who defines the "separation change" for each sentence; Sanford *et al.* (1994) who use a variant of this; de Haan and Schils (1993b) who examine discrepancies between the observed and expected cusum charts; and finally Hilton and Holmes (1993) who investigate the efficacy of a variant of the cusum test based on weighted cusums and which allows for more objective and rigourous testing.

4.1. Canter — An Evaluation of the "Cusum" stylistic analysis of confessions

Canter (1992) states that no published accounts can be found to indicate what the psychological or linguistic processes are that generate the "habit" which proponents of cusum claim to be consistent; indeed no aspect of human behaviour that has been studied to date reveals such high levels of consistencies. He questions whether the habit is anything other than a random property of the arithmetic of language. To the extent that two and three letter words are random constituents of any language, there will be a consistent relationship between their frequency and the number of words in the sentence as a whole. Variations from that consistency could be random artifacts of no substantive significance.

Canter proceeds to analyse more than one hundred cusum charts using Spearman's rank correlation coefficient, defined as

$$\rho = 1 - \left\{ \frac{6}{n(n^2-1)} \sum_{i=1}^{n} |d_i|^2 \right\}, \tag{6}$$

where d_i is the difference between the ranked sentence-length cusum values and ranked habit cusum values and n is the number of sentences as before. A cusum plot from texts by a single author ought to have a value of the coefficient, ρ, close to 1, while text of multiple authorship should yield a ρ value further away from 1. Canter proposes 0.9 as a cut off point, a value often used to assess the reliability of psychometric tests. Canter then carries out a series of tests using ρ to examine the reliability of cusums and found that 48 % of single-author texts have $\rho < 0.9$, implying multiple authorship. The texts were then split into sections and recombined with samples selected randomly from different authors. In this case 65 % of the texts had $\rho > 0.9$, mistakenly indicating single authorship. To assess this procedure with texts that courts were likely to come into contact with, a similar analysis of uncontested statements or personal accounts was carried out. Here 29 % of the texts by a single author were misclassified as being of multiple authorship, the case most likely to be examined by a court. The ρ statistic does not, therefore, reliably distinguish between single and multiple author texts, although more research could be done here using cut-off values other than 0.9.

Robertson and Vignaux (1993) perform likelihood ratio tests on these results to confirm that this particular method of stylometry would be worthless as evidence in determining whether samples had been written by one or more than one person. The likelihood ratio approach to the evaluation of evidence

in the law has been discussed at the RSS Conference, Sheffield, 1992 [see Aitken (1993)].

Canter then moves on to examine the objectivity of purely visual tests, as advocated by Morton. A research assistant, sympathetic to Morton's claims, was asked to assess the cusum plots generated above without reference to the text or knowing the origin of the material. The percentages judged to be of mixed authorship were virtually identical (62 % – 68 %) for each set of material, whether single authorship, mixed authorship, written or spoken. When visual inspection is the sole mode of determination, therefore, there is actually an increased risk of mistakenly assuming mixed authorship.

Canter concludes that his studies do not support the hypothesis that variations in cusum plots are a consequence of mixed authorship. He queries whether the variations are of substantive psychological significance at all or whether they simply reflect the individual psychology of the person making the utterances, the cognitive demands of the material or the social processes within which the utterances are made.

4.2. Hardcastle — Forensic Linguistics: an assessment of the cusum method for the determination of authorship

In an attempt to aid the visual interpretation of cusum charts, Hardcastle (1993) plots a function of the vertical separations between the sentence-length plot and the habit plot. He defines the "separation change" for each sentence as the observed habit words minus the expected habit words, where the number of expected habit words is equal to the sentence length multiplied by the average number of habit words per sentence, divided by the average sentence length. A histogram may then be constructed of the differences between observed and expected values throughout the text, a large difference (positive or negative) indicating that the sentence departs markedly from the average rate of use of the habit words.

On comparing these histograms with their parent cusum charts for several specimen texts, Hardcastle finds that a considerable variation in habit rate occurs from one sentence to the next within a single-author text, this range being larger for spoken material as compared to written material. It is possible, therefore, that if a cluster of high or low values occurs within a questioned text, it could be erroneously interpreted as a section of text of different authorship to the rest. Hardcastle's histograms also illustrate an important weakness of

the cusum method, namely that when a deviation occurs between the two parent cusum lines, the cumulative nature of the chart means that this separation is "carried forward" to the following part of the chart.

Hardcastle points out how Morton, whilst stressing that it is not the detailed profiles of the cusum lines that are important but the separations between them, in using transparencies to allow one cusum line to be rotated or translated with respect to the other paradoxically ignores the separations and compares, instead, details of the line shapes. Hardcastle concludes:

> "It is clear that, for texts of forensic interest at least, the cusum method in its present form cannot be regarded as objective or reliable."

4.3. Sanford *et al.* — A critical examination of assumptions underlying the cusum technique of forensic linguistics

Sanford *et al.* (1994) use a similar technique to that of Hardcastle (1993) to measure the differences between cusum charts. As the second part of their paper they construct sets of mixed texts with 15 sentences from one source with an insert of 10 sentences from another source starting at sentence 5, making a 25 sentence sample. Four such sets are made up, the first with the same author and a small difference in ratio of the habit, the second with the same author and a large difference in habit ratio and two similar sets with different authors.

Sanford *et al.* (1994) examine the goodness of fit of their cusum plots by calculating the sum of the deviations of pairs of points in the plot in a similar manner to Hardcastle (1993). The resulting analysis of variance shows that there is no statistically significant difference between charts derived from a single author and mixed authors. There is, however, a significant difference between the ratio differences, there is less deviation between charts from sources with similar ratios than from sources with distant ratios. This directly contradicts Morton's claims about the technique; deviations in cusum charts are not indicators of authorship, merely of a change in habit rate.

Sanford *et al.* (1994) conclude that the cusum method is "open to unknown subjective distortions", "without any reliable foundation" and that

> "until the method receives more adequate support, it should not be entertained as a forensic technique."

4.4. De Haan and Schils — The QSum plot exposed

De Haan and Schils (1993b) suggest that, in order to facilitate the interpretation of his cusum plots, Morton should plot one graph only, a graph representing the "distance" between values D1 (the discrepancy between the observed and the expected cumulative sum of the habit) and D2 (the discrepancy between the observed and the expected cumulative sum of the sentence lengths). They argue that if the text consists of two sub-fragments with different underlying habits, then this graph will, in the ideal case, reverse its slope at the join of the two fragments. In reality, habits in both homogeneous and heterogeneous texts are prone to sampling error and it can become quite difficult to determine the location of the "bend" in the suggested graph. De Haan and Schils illustrate this by a series of simulations in which a sample of sentence lengths containing a particular rate of occurrence of a habit are drawn, joined to another sample of sentence lengths with a different habit rate, the whole being subjected to a normally distributed error term. They find that the larger the error term, the smaller is the success rate in localizing the join between the two sections of text. Using both binomial and empirically assessed error levels, on only 13 and 15 occasions, respectively, out of a series of 100 simulations, was the join correctly localized. These findings confirm that, given the natural variation in rates of occurrence of habit words throughout a text, identifying what is thought to be a change in authorship is a highly subjective procedure.

4.5. Bissel, Hilton and Holmes — An assessment of cumulative sum charts for authorship attribution

Whilst the cusum test appears at times to produce remarkable results using very small samples of text, the subjectivity of the technique in evaluating differences between pairs of superimposed cumulative sum charts does give cause for concern. The search for an objective method of examining cusum charts motivated the work done by Hilton and Holmes (1993).

A. F. Bissell, in private correspondence, has suggested the use of *weighted* cumulative sums for the analysis of features such as short-word frequency and use of nouns or other parts of speech. The advantage of the weighted cumulative sum is that it uses sentence length as an explanatory variable, thus removing its influence from the slope of the cumulative sum chart and leaving one able to concentrate on changes in proportion or rate of use of some subsets of words, without the problem of lining up superimposed

charts. There is no obvious way of testing the significance of discrepancies of superimposed cumulative sum plots, but with weighted cumulative sums we may conduct formal testing procedures employing well established techniques. To investigate differences between two texts, one can estimate average and variation within each text and then carry out quasi-t tests between the texts. An expository paper on the technique has been written by Bissell (1990).

The values of the weighted cumulative sum, s_i, are calculated according to the formula

$$s_i = \sum_{r=1}^{i} (x_r - \bar{p} w_r) \tag{7}$$

where x_r is the number of times the "habit" in consideration occurs in sentence r, w_r is the length, in words, of sentence r, and $\bar{p}$ is the proportion of words in the sample which signify an occurrence of the habit. The quantity $\bar{p}$ is called the "weight", and is calculated according to the formula

$$\bar{p} = \frac{\sum_{r=1}^{n} x_r}{\sum_{r=1}^{n} w_r} \tag{8}$$

for a text n sentences in length. Thus, for each sentence, the measure s_i is the cumulative sum of the difference between the observed number of habit occurrences and the "expected" number of occurrences. When plotting a weighted cusum, s_i is plotted on the vertical axis, and the cumulative sentence length divided by the average sentence length,

$$\frac{\sum_{r=1}^{i} w_r}{\bar{w}} \tag{9}$$

is plotted on the horizontal axis.

As an illustration, consider Table 1 in which columns two and three give the length of sentence (w) in words, and the number of nouns in that sentence (x), respectively, for a sample of twenty sentences from Author A followed by a sample of twenty-one sentences from Author B. In this example,

$$\bar{p} = \frac{147}{661} = 0.2224 . \tag{10}$$

The weighted cumulative sum, s, is given in column four of Table 1. Column five lists the cumulative sentence length divided by the average sentence length, where the average sentence length in this example is 16 words. Figure 16 shows the weighted cusum plot for this example.

The plot in Figure 16, however, is no longer a subjective visual indicator of a difference in slopes (and hence a difference in authorship) between Text A and Text B. We can determine if there are significant differences in the rates

Tableau 1

Computing the weighted cumulative sum of noun usage within a text sample

Sentence No.	Sentence		Weighted	
	Length (w)	Nouns (x)	Cusum (s)	$\sum \frac{w}{16}$
1	16	4	0.44	1.00
2	6	2	1.11	1.38
3	14	6	3.99	2.25
4	15	3	3.66	3.19
5	6	0	2.32	3.56
6	14	5	4.21	4.44
7	6	1	3.88	4.81
8	11	1	2.43	5.50
9	11	3	2.98	6.19
10	23	5	2.87	7.63
11	15	4	3.53	8.56
12	8	3	4.75	9.06
13	9	1	3.75	9.63
14	11	4	5.30	10.31
15	25	3	2.74	11.19
16	27	4	0.74	13.56
17	26	5	−0.04	15.19
18	3	0	−0.71	15.38
19	9	2	−0.71	15.94
20	15	3	−1.05	16.88
21	28	7	−0.28	18.63
22	10	3	0.50	19.25
23	20	5	1.05	20.50
24	18	4	1.05	21.63
25	24	6	1.71	23.13
26	17	4	1.93	24.19
27	9	2	1.93	24.75
28	6	2	2.60	25.13
29	15	4	3.26	26.06
30	13	3	3.37	26.88
31	11	2	2.92	27.56
32	47	11	3.47	30.50
33	27	2	−0.54	32.19
34	11	2	−0.98	32.88
35	20	6	0.57	34.13
36	34	5	−1.99	36.25
37	9	1	−2.99	36.81
38	14	6	−0.11	37.69
39	11	2	−0.55	38.38
40	36	7	−1.56	40.63
41	11	4	0.00	41.31
Total	661	147		

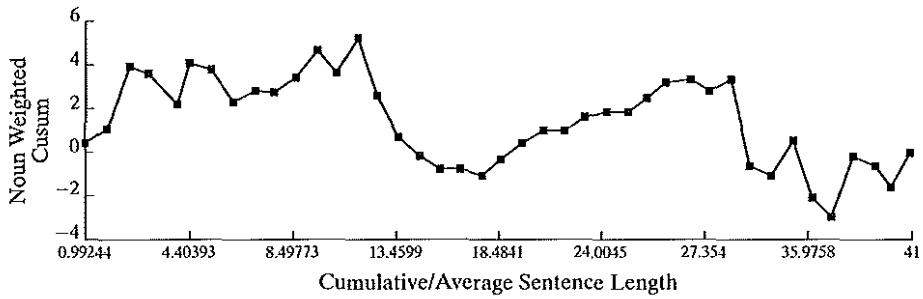

Fig. 16.– Weighted cumulative sum plot of data in Table 1

of occurrence of the observed habit by conducting a formal testing procedure using the quasi-t test. Text A and Text B are treated as separate samples and the test statistic

$$t = \frac{\text{Weight } A - \text{Weight } B}{\sqrt{\frac{\text{Variance } A}{\text{Total Words } A} + \frac{\text{Variance } B}{\text{Total Words } B}}} \tag{11}$$

is calculated. The problem of calculating the variance of data with varying element sizes (sentence lengths) has been covered by Bissell (1991). Bissell advocates using weighted squared successive differences to estimate σ^2, the underlying variance. This gives:

$$\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n-1}\left\{\frac{\left(\frac{x_i}{w_i} - \frac{x_{i+1}}{w_{i+1}}\right)^2}{\frac{1}{w_i} + \frac{1}{w_{i+1}}}\right\}. \tag{12}$$

The formula for t can now be written as

$$t = \left|\frac{\overline{p_A} - \overline{p_B}}{\sqrt{\frac{(\widehat{\sigma_A})^2}{\sum w_A} + \frac{(\widehat{\sigma_B})^2}{\sum w_B}}}\right| \tag{13}$$

The t-value computed with the above formula is a measure of the evidence against the null hypothesis that the frequency of usage of the habit under consideration is the same in Text A and Text B. The higher the t-value, the more evidence against the hypothesis. The number of degrees of freedom, ν, for the t-test is obtained using

$$\nu = \frac{2}{3}(n_A + n_B - 2) \tag{14}$$

where n_A and n_B are the number of sentences in Texts A and B respectively and the multiplication by $\frac{2}{3}$ is to allow for the correlation in the variance, [Bissell and Williamson (1988)].

Let us now return our attention to the example data presented in Table 1. Using equations 8 and 12, we calculate the weights for Text A and Text B as 0.2185 and 0.2251 respectively, and their standard deviations as 0.3927 and 0.3708, respectively. There are 270 words in Text A and 391 words in Text B. Substituting these values into the formula for t yields

$$t = \left| \frac{0.2185 - 0.2251}{\sqrt{\frac{(0.3927)^2}{270} + \frac{(0.3708)^2}{391}}} \right| \tag{15}$$

$$= 0.22 \tag{16}$$

For our example, which has 26 degrees of freedom, the critical value of t at the 5 % level of significance is 2.056. Clearly, the t value of 0.22 is less than the critical t-value and we have found no evidence of any difference between Author A and Author B regarding the rate of noun usage in the textual passages sampled.

The weighted cumulative sum technique described in this section is based on the same assumptions as the cusum test, with both tests comparing how closely the usage of a particular class of words in one text-sample matches the usage in another sample. The weighted cusum test is more powerful than the cusum test in that it enables us to present sequential data with a variable element size structure and to conduct a formal testing procedure on segments of the data.

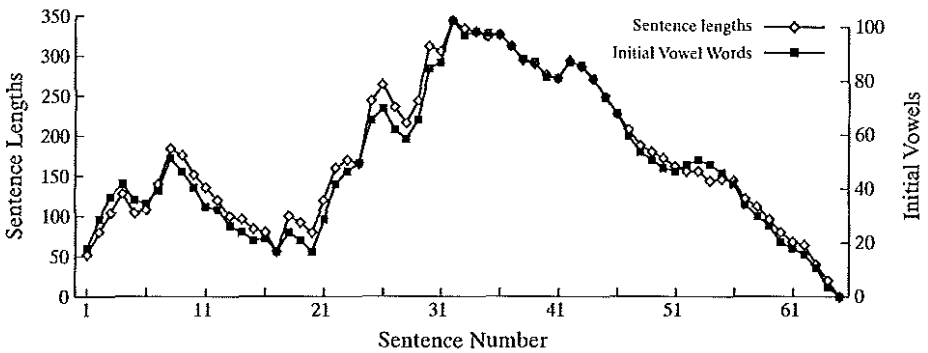

Fig. 17.– Fleming's *Chitty-Chitty-Bang-Bang* and *From Russia With Love*

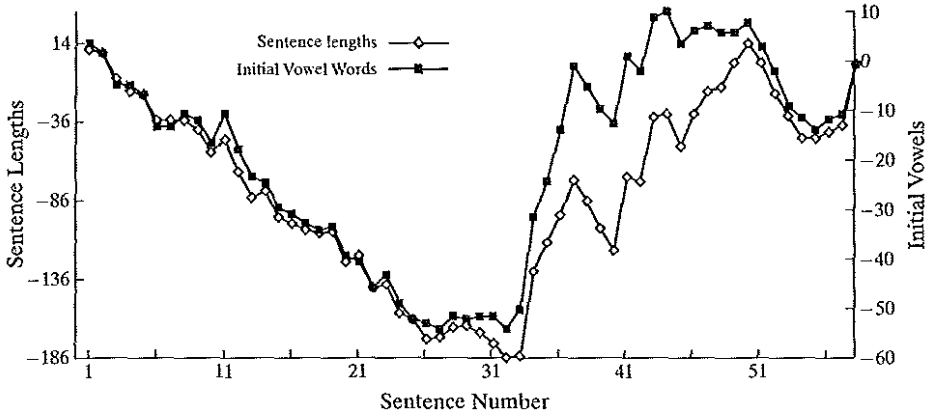

Fig. 18.– Amis' *Colonel Sun* and Fleming's *Octopussy*

Hilton and Holmes (1993) assess the efficacy of weighted cusums for authorship attribution. The texts chosen for their experiments covered different genres and time periods, and were tested for several different "habits". The first test group contained four twentieth-century books: *From Russia With Love*, *Octopussy* and *Chitty-Chitty-Bang-Bang* by Ian Fleming, and *Colonel Sun* by Kingsley Amis, the latter being written in imitation of Fleming's "James Bond" novels. In each test, approximately 30 sentences from each of two of these books were concatenated. Figures 17 to 19 show Morton's cusum plots for three of these tests. Visual judgement attributed "correct" results for Morton's cusums in five out of eighteen tests, correctness being defined as whether or not a significant discrepancy was observed near the point of concatenation. The weighted cusum test, using t-values, performed marginally better with eight correct cases. Four out of six cases which were indeterminate using Morton's visual method, were correctly decided by the weighted cusum test.

The second test group covered by Hilton and Holmes concerned textual samples taken from *The Federalist Papers*. Weighted cusums correctly determined whether or not there was an authorship shift at the point of concatenation on fourteen out of thirty cases, initial-vowel words performing best with 70 % correct classifications compared with a 20 % success rate for nouns.

The final test group contained five novels written by Jane Austen. The t-values for weighted cusum tests of Austen versus the Other Lady in *Sanditon* were 0.36, 0.96 and 0.31 for initial-vowel words, two and three-letter words and nouns respectively, none of these being large enough to indicate dual

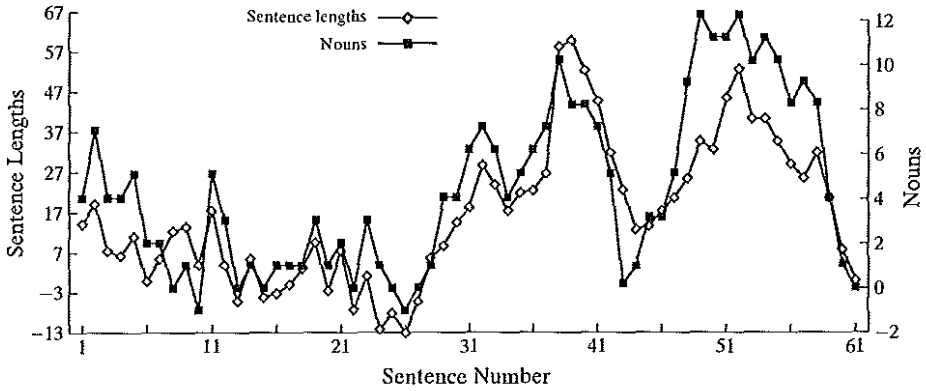

Fig. 19.– Amis' *Colonel Sun* and Fleming's *From Russia With Love*

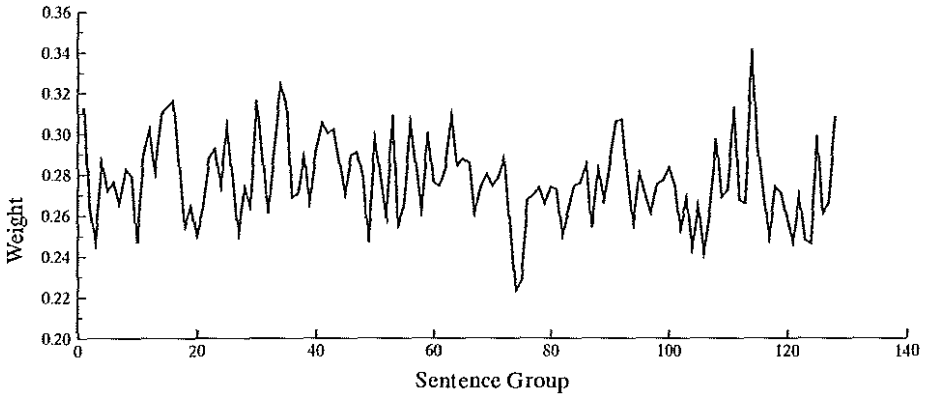

Fig. 20.– Frequency of initial-vowel words in Austen's *Mansfield Park*

authorship. The text of four of Austen's full novels were available in machine-readable form allowing Hilton and Holmes to conduct weighted cusum tests on the complete text of each novel. Using initial-vowel words, only half of the results correctly identified single authorship when two novels were concatenated. This is a surprising result since if an author does follow a habit consistently, then the larger a sample of the writer's text one examines, the more accurately one should be able to quantify the frequency of the habit. Figure 20 confirms that the frequency of initial-vowel word usage does indeed vary throughout Austen's *Mansfield Park*.

In the light of their results, Hilton and Holmes conclude that weighted cusums perform marginally better than Morton's visual method, but the lack

of stability for different sample sizes casts serious doubt on the claim that authors are internally consistent with habits such as usage of initial-vowel words. These cumulative sum techniques, therefore, do not appear to give consistently reliable results in studies of authorship.

5. Conclusions

Morton's cusum technique has provoked much controversy since its introduction in 1990. Critics have said that it is subjective, lacking in scientific rigour and relies more on the interpreter's knowledge of the text rather than on the ill-defined results produced. Advocates of the technique, however, brush these comments aside claiming that it is robust and that any detailed statistical analysis would be above the heads of a judge, lawyer or layman.

In this paper we have presented the technique and the criticisms directed against various aspects of it. We have shown that the method is riddled with poor practice. The method itself has only been presented in technical reports and legal journals, and not in any refereed statistical or linguistic journals. It is a visual technique and hence subjective, no statistical measures are entertained. The use of the technique by its proposers also leaves much to be desired with changes in the scale of graphs, editing of the texts concerned and criteria for matching cusums plots that change with each report.

In an attempt to evaluate the cusum technique various statistical techniques have been used, from Spearman's ρ to analysis of variance. Not one of the papers that we review gives any support to the technique. Another method, that of weighted cusums, was presented and extended using multiple comparisons. Although this objective technique addresses some of the criticisms directed against Morton's cusums, it only performs marginally better than the visual method.

It is clear from what has been presented in this paper that the cusum method should not be seen as a test of authorship.

Bibliography

AITKEN (C.G.G.): 1993, "Statistics and the Law: Conference report from RSS conference, Sheffield, September 1992". *Journal of the Royal Statistical Society Series A*, 156(2): 301–304.

BEE (R.E.): 1971, "Statistical methods in the study of the Masoretic Text of the Old Testament". *Journal of the Royal Statistical Society Series A*, 134(4): 611–622.

BEE (R.E.): 1972, A statistical study of the Sinai Pericope. *Journal of the Royal Statistical Society Series A*, 135(3): 406–421.

BISSELL (A.F.): 1990, Weighted cusums — method and applications. *Total Quality Management*, 1(3): 391–402.

BISSELL (A.F.): 1991, Estimating variation from data with varying element sizes. *Journal of Applied Statistics*, 18(2): 287–295.

BISSELL (A.F.) and WILLIAMSON (R.J.): 1988, Successive difference tests — theory and interpretation. *Journal of Applied Statistics*, 15: 305–323.

CAMPBELL (D.): 1992, Writings on the wall. *The Guardian*, p. 25.

CANTER (D.): 1992, An evaluation of the "Cusum" stylistic analysis of confessions. *Expert Evidence*, 1(3): 93–99.

DAVIES (H.E.): 1990, Uncertain authorship and the Cusum technique. Master's thesis, University of Wales, Swansea.

DE HAAN (P.) and SCHILS (E.): (1993a), Characteristics of sentence length in running text. *Literary and Linguistic Computing*, 8(1): 20–26.

FARRINGDON (M.G.): 1992, The case of the Philosophy Department documents: Two anonymous documents analysed and compared with a letter by Cumulative Sum stylometry.

FARRINGDON (M.G. and MORTON (A.Q.): 1990, Fielding and the Federalist. Technical Report CSC 90/R6, University of Glasgow.

HALLIDAY (M.A.K.): 1990, *Spoken and written language*. Oxford University Press.

HARDCASTLE (R.A.): 1993, Forensic linguistics: an assessment of the CUSUM method for the determination of authorship. *Journal of the Forensic Science Society*, 33(2): 95–106.

HARRISON (D.): 1993, Royal tapes are fakes, says language expert. *Observer*, pages 1–2.

HILTON (M.L.) and HOLMES (D.I.): 1993, An assessment of cumulative sum charts for authorship attribution. *Literary and Linguistic Computing*, 8(2): 73–80.

HOLMES (D.I.): 1992, A stylometry analysis of Mormon scripture and related texts. *Journal of the Royal Statistical Society Series A*, 155(1): 91–120.

JAMIESON (F.J.) and AITKEN (C.G.G.): 1993, A survey of approaches in authorship studies. University of Edinburgh, Department of Computer Science 4th Year Report.

MATTHEWS (R.): 1993a, Harsh words for verbal fingerprint. *Sunday Telegraph*.

MATTHEWS (R.): 1993b, Linguistics on trial. *New Scientist*, (1887): 12–13.

MICHAELSON (S.) MORTON (A.Q.) and WAKE (W.C.): 1978, Sentence length in Homer and hexameter verse. *Association for Literary and Linguistic Computing Bulletin*, 2.

MORTON (A.Q.): 1991, Proper words in proper places. Technical Report 91/R18, University of Glasgow, Computing Science Department.

MORTON (A.Q.) and FARRINGDON (M.G.): 1992, Identifying utterance. *Expert Evidence*, 1(3): 84–92.

MORTON (A.Q.) and MICHAELSON (S.): 1990, The Qsum plot. Technical Report CSR-3-90, University of Edinburgh.

ROBERTSON (B.W.N.) and VIGNAUX (G.A.): 1993, Expert evidence — probability — Bayes' theorem — need for court to decide prior probabilities. *Expert Evidence*, 2(1): 47–48.

SANFORD (A.J.) AKED (J.F.) MOXEY (L.M.) and MULLIN (J.): 1994, A critical examination of assumptions underlying the cusum technique of forensic linguistics. *Forensic Linguistics*, 1(2): 151–167.

TAVES (E.H.): 1991, *This is the Place: Brigham Young and the New Zion*, chapter 17: A New Stylometry, pages 265–276. Prometheus books, Buffalo, NY.