Greek-language authors, he found that the variability in sentence lengths between the various epistles was greater than that found in any corpus by a single author. He suggested that Romans I and II, Corinthians and Galatians could be attributed to Paul, while the other Epistles were the work of up to six different authors. Kenny (1982) reports that a weakness in Morton's argument is that his methods also reveal anomalies in II Corinthians and Romans itself.

Morton (1965) studied the relative frequency of the word *kai* (the commonest word, meaning *and*) in the Greek epistles of the New Testament generally attributed to St Paul. Table 5.5 shows Morton's results for 11 of the epistles, according to the use of *kai* as a percentage of the total number of words, and the percentage of sentences containing *kai*. The data appears to fall into two groups, with Romans, Galatians and Corinthians I and II producing similar results as if written by one author, and the other epistles producing significantly different results as if written by another author or authors.

	***kai* as a percentage of all words**	**Percentage of sentences containing *kai***
Romans	3.86	33.7
I Corinthians	4.12	41.6
II Corinthians	4.41	40.7
Galatians	3.22	29.3
Ephesians	5.67	67.0
Philippians	6.56	51.7
Colossians	6.27	51.8
Thessalonians I	6.73	58.0
Timothy I	5.72	54.2
Timothy II	5.41	48.9
Hebrews	5.16	49.1

Table 5.5 The usage of *kai* in epistles usually attributed to St Paul

These studies of sentence length and Greek function words such as *kai* showed, according to Morton, that St Paul wrote only four of the New Testament epistles traditionally attributed to him, since the others show less consistency of style. However, when Ellison (1965) used these same attributes on English texts, they showed that James Joyce's *Ulysses* and Morton's own essays must each have been written by several authors. Other criticisms of Morton's methods were that he encoded his Greek texts with modern punctuation and also that sentence length varies considerably within random samples of text. Morton later suggested new stylistic criteria such as the part of speech of words coming last in a Greek sentence, the number of words between successive occurrences of *kai*, the relative occurrence of genitive and

non-genitive forms of the Greek pronoun *autos* (*he*), and pairs of words in English which collocate such as *of course* and *as if* (Oakman 1980).

2.5 Ellegård's work on the Junius letters

The Junius letters are a series of political pamphlets written between 1769 and 1772 under the pseudonym Junius. The authorship of the letters has been attributed at various times to no fewer than 40 different people. Ellegård (1962) compared the anonymous Junius letters and their 157,000-word context with 231,000 words of text by Sir Philip Francis (the most probable contender for authorship of the Junius letters) and other 18th-century English writers. He found that Yule's sentence-length test was not sensitive enough to discriminate between many authors. The K characteristic could not be used since it requires large text samples of at least 10,000 words and some Junius samples are less than 2000 words.

Ellegård presented a list of 458 words and expressions (mainly content words) designated either 'plus' (occurring more frequently in the Junius letters than in work by other authors) or 'minus' (occurring less frequently in the Junius letters than in work by other authors) indicators of style. He calculated their occurrence in all the sample texts. The ratio of usage of a word in samples of an author's text compared with its usage in a representative sample of contemporary texts is called the **distinctiveness ratio**. The plus words have a distinctiveness ratio above 1, while the minus words have a ratio below 1. For example, the relative frequency of the word *uniform* (when used as an adjective) was 0.000280 in the sample of Junius texts, but only 0.000065 in the comparison sample of one million words. The distinctiveness ratio of *uniform* is then 0.000280/0.000065 which is about 4.3, showing that *uniform* is a plus word. In order to produce a reliable 'testing list' of plus and minus words, a text of at least 100,000 words is required. However, once this list has been produced, it can be used to test much smaller samples. Ellegård's study of 'plus' and 'minus' words showed that Sir Philip Francis was the most likely author of the Junius letters.

As Ellegård did not use a computer for counting the words, he had to rely on his own intuition as to which ones occurred more or less frequently than expected. The study would have been more comprehensive if he had used a computer for the counting and to have counted all words, not just those he specifically noticed (Hockey 1980).

Ellegård also identified about 50 pairs or triplets of approximate synonyms such as *on/upon*, *kind/sort*, *and/also*, *since/because* and *scarcely/hardly*. Their patterns of usage in the Junius letters showed strong correspondence with the textual characteristics favoured by the author Sir Philip Francis, while the other authors were clearly distinct.

Austin (1966) also found lists of 'plus' words which were able to discriminate between the work of Robert Greene and Henry Chettle, both possible authors of *The Groatsworth of Wit* written in 1592. Once all spelling and inflectional

variants of each word had been merged, 'plus' words were taken to be those which fulfilled the following criteria:

1. the word must occur at least ten times in one author
2. the frequency per 1000 words in one author must exceed the corresponding frequency in the other by 1.5 – the differential ratio
3. the ratio of variation within the body of works of one of the authors must be lower than the differential ratio.

Fifty words were found which fulfilled all three criteria. Austin's work demonstrates the three stages of automated authorship studies. Firstly, the computer can be used to look in large bodies of text for some stylistic feature with discriminating power; secondly, the disputed texts are searched for the presence or absence of this feature; and finally, a statistical analysis is performed to test the validity of the findings (Oakman 1980).

2.6 The Federalist papers

The so-called Federalist papers were published in newspapers in 1787–8 to persuade the population of New York state to ratify the new American Constitution. Published under the pseudonym Publius, their three authors were James Madison, John Jay and Alexander Hamilton. In 1804, two days before his death in a duel, Hamilton produced a list of which essays were written by which author. He left this list in the home of a friend called Egbert Benson. In 1818 Madison claimed that some of the essays in the Benson list attributed to Hamilton were in fact written by himself. Altogether, it was agreed by both Hamilton and Madison as well as later historians that Jay wrote five of the essays, 43 were written by Hamilton and 14 by Madison. Another 12 were disputed, and three were written jointly by Hamilton and Madison (Francis 1966). In Sections 2.6.1 and 2.6.2 we will look at two contrasting approaches for determining the authorship of the disputed papers.

2.6.1 The Bayesian approach of Mosteller and Wallace

The stylistic features of the disputed papers could be compared with the features in the papers of known authorship. Unlike the case of the Junius letters which might potentially have been written by any of a large number of authors, in the case of the disputed authorship of the Federalist papers, there are only two candidates, Hamilton and Madison. Mosteller and Wallace (1963) used Bayesian statistics, described in Chapter 1, Section 6, to determine which of the disputed papers were written by which author, hoping that these methods would lead to the solution of other authorship problems. Like Ellegård, Mosteller and Wallace combined historical evidence with statistical computation.

The problem of deciding which variables to use to discriminate between the writing styles of Hamilton and Madison is difficult, because both adopted a writing style known as **Addisonian** which was very popular in their day. Since the two authors may consciously have tried to imitate each other's style, stylistic features which can be easily imitated – sentence length, for example – had to be discounted from the analysis. In this respect, variations in the use of high-frequency words might be a reliable criterion, since authors probably would not be conscious of using these. It is also desirable that variables should not vary with context (for example, sentence length in legal documents is greater than in most other texts), but must have consistent rates over a variety of topics. The occurrence of function words was examined since these tend to be non-contextual, except for personal pronouns and auxiliary verbs. The styles of Madison and Hamilton were also found to vary significantly in the frequency of use of various function words, such as *by* (more characteristic of Hamilton) and *to* (more characteristic of Madison). Some low-frequency words were also used since they tended to occur almost exclusively in the work of one author. For example, in a sample of 18 Hamilton and 14 Madison papers, *enough* appeared in 14 Hamilton papers and no Madison papers, and *whilst* appeared in no Hamilton but 13 Madison papers. Altogether, 28 discriminating terms were found.

Two mathematical models exist which describe the frequency distribution of individual words across equal-sized portions of text. In the Poisson model, the occurrence of a given word is independent of the previous occurrence of that word – it depends only on its overall rate of occurrence. However, a writer might consciously avoid repetition of words on one hand, and on the other may repeat a word for emphasis or parallelism (Francis 1966). The tendency for the same word to appear in clusters is taken into account by the alternative negative binomial model (see Agresti 1990). The Poisson distribution is described by a formula which, if given the average rate of occurrence of a word in a text of certain length in words (a single parameter), will give the proportions of text sections of that length which have none, one, two, etc., occurrences of the word. The Poisson formula takes the form

$$p_n = \frac{\lambda^n \times e^{-\lambda}}{n!}$$

where e is the constant, and λ is the average number of times the word occurs per section of text. p_n is the proportion of text sections which have n occurrences of the word. $n!$ means n multiplied by $(n-1)$, then multiplied by $(n-2)$ and so on until we reach 1. Thus $3! = 3 \times 2 \times 1$. By convention, $0! = 1$. e to the power x is often called the exponential (exp) of x, and may be calculated using a typical scientific calculator. For example, if the average occurrence of a given word per section of text (λ) = 0.1, then the proportion of text sections with no occurrences at all of the word (p_0) is $e^{-0.1} = 0.905$, the proportion of text sections with just one occurrence (p_1) is $0.1 \times e^{-0.1} = 0.0905$,

$$p_2 = \frac{(0.1)^2 \times e^{-0.1}}{2} = 0.0045, \text{ and}$$

$$p_3 = \frac{(0.1)^3 \times e^{-0.1}}{6} = 0.0001.$$

The negative binomial formula requires two parameters: the average rate of occurrence of a word and its tendency to cluster (non-Poissonness parameter). The negative binomial distribution fitted the observed data better than the Poisson distribution.

Francis (1966) gives the following example of the use of Bayesian statistics. Imagine the word *also* is used 0.25 times per 1000 words by Hamilton and 0.5 times per 1000 words by Madison, and that both authors follow the Poisson model. Imagine too that the word *also* occurs four times in a 2000-word paper. The Poisson probabilities for a word from a 2000-word paper, Hamilton's rate per 1000 words being 0.25 and Madison's 0.5, for four occurrences are 0.00158 for Hamilton and 0.0153 for Madison. The calculation for Hamilton was performed as follows: since Hamilton used the word *also* 0.25 times per 1000 words, this means that he used *also* at a rate of 0.5 times per 2000 words (the length of the text of unknown authorship). This rate of 0.5 becomes λ in the Poisson equation. Since we are interested in the probability of four occurrences of *also* in a 2000-word text, n is 4, and the bottom line of the equation is 4! = 4 x 3 x 2 x 1 = 24. This gives

$$p_4 = \frac{(0.5)^4 \times e^{-0.5}}{24} = \frac{0.0625 \times 0.6065}{24} = 0.00158$$

Thus, it is more likely that Madison wrote the paper, with a likelihood ratio of 0.0153/0.00158 giving odds of about 10 to 1. If we had had a prior opinion that Madison was the more likely author with odds of 3 to 1, this new evidence would be multiplied by the prior belief to yield posterior odds of 30 to 1. The next piece of evidence that we will examine is that the word *an* appears seven times in the disputed paper. Using tables of the Poisson distribution we find that the likelihood ratio for this eventuality is 0.0437/0.111 which is about 3/8. The old posterior odds become the new prior odds, and these are multiplied by this new likelihood ratio to give new posterior odds of 80 to 1. This process continues by considering the usage of a range of words whose discriminating power is high.

One of the assumptions made in this simplified analysis given by Francis is that the average rates of usage of each word by each author (the parameters of the model) were known. However, the true rates are unknown. We can only count the rates of occurrence in each word in **available texts** known to be written by each author, but we cannot possibly obtain all texts ever written by each author, and some existing texts we would exclude as their authorship remains controversial. The parameters must be estimated from prior information

(for example, existing studies of word rates) and sample data (94,000 words of text by Hamilton and 114,000 words written by Madison).

Since the exact values of the authors' word-usage rates are not known, Francis maintains that it is better to express the knowledge that we do have in the form of probability distributions rather than point estimates such as 0.25 and 0.5. Thus, rather than stating that Hamilton's rate for a given word is 0.25, we could state, using hypothetical figures, that there is a probability of 1/20 that his rate is less than 0.15, a probability of 1/10 that it is between 0.15 and 0.20 and so on. Once again, Bayes's theorem is employed to combine the information contained in the prior word-frequency distribution with the information contained in the samples of known works of the two authors about relative rates of word usage summarised as a likelihood ratio. Because of the imprecision in the prior information, Mosteller and Wallace chose several possible distributions to be sure that at least one fitted the true prior distribution accurately, and carried out their analysis using each of these in turn.

Thirty words were eventually chosen for the main study, since a pilot study showed them to be good discriminators between Madison and Hamilton. Some words were discarded because they had different rates of usage when used by Madison within and outside the Federalist corpus. The final list is given in Table 5.6.

Group B3A	upon
Group B3B	also an by of on there this to
Group B3G	although both enough while whilst always though
Group B3E	commonly consequently considerable(ly) according apt
Group B3Z	direction innovation(s) language vigor(ous) kind matter(s) particularly probability work(s)

Table 5.6 Final words and word groups used by Mosteller and Wallace

The results were reported not in odds but in log odds, which are the natural logarithms of the odds, to restrict the range of results to more manageable values. (Natural logarithms or $\log_e$ (sometimes written 'ln') can also be calculated using a typical scientific calculator.) Positive values indicate a verdict in favour of Hamilton, while negative values indicate a verdict in favour of Madison. The method of Mosteller and Wallace was checked by applying it to 11 papers known to be written by each author (eight taken from the Federalist source and three taken from external sources). The resulting data, when assuming a negative binomial distribution, reveals that every Hamilton paper has positive log odds and every Madison paper negative log odds. This is evidence that their method is accurate. When the method was employed to obtain log odds for the three joint and 12 disputed papers, for every prior distribution tested the resulting log odds were greatly in favour of Madison being their author.

Oakman (1980) notes that Mosteller and Wallace's expectation that the distribution of function words may be a general signifier of authorship has been thrown into doubt by Damerau (1975), who found a great disparity in the use of function words in samples of *Vanity Fair* and three American novels. Thus a particular word list worked well for the Federalist papers, but the same list cannot necessarily be used for other authorship studies.

2.6.2 Biological analogy and the Federalist papers: recent developments

Tweedie, Singh and Holmes (1994) show that statistical methods of authorship attribution can be used in conjunction with a neural network to provide an effective classification tool. Previous studies of the Federalist papers had used function words, and in their preliminary experiments Tweedie, Singh and Holmes also chose common words on the basis of their ability to discriminate between the work of Hamilton and Madison. These words were *an, any, can, do, every, from, his, may, on, there* and *upon*. The number of occurrences per thousand words of each of these words was found using the Oxford Concordance Program, described in Chapter 4, Section 2.4.2. These values were converted to z scores as described in Chapter 1, Section 2.4, so that each word had a rate that was normally distributed with a mean of 0 and a variance of 1. This ensures that each word contributes equally to the neural network training process.

Neural networks simulate the way neurons interact in the human body. Each neuron has a body, and receives stimuli from other neurons or the environment. The human neuron is either at rest or firing. It fires whenever the sum of the stimuli it receives exceeds a certain threshold, and when it fires, other neurons are stimulated or an action such as a movement is taken. Tweedie, Singh and Holmes used an array of computer-simulated neurons called a multilayer perceptron, which consisted of a so-called input layer of 11 neurons, each of which was stimulated to a degree which depended on the occurrence of one of the 11 marker words in a given text. These neurons were connected to a middle or 'hidden' layer of three neurons, and these in turn were connected to an 'output' layer of two neurons. One of these would fire if the network thought the text was written by Hamilton, and the other would fire if the network thought the text was written by Madison. The number of neurons in the input and output layer are clearly related to the number of possibilities within the task at hand, but there are no hard and fast rules for deciding on the most effective number of neurons in the hidden layer. A small part of this neural network is given in Figure 5.2, where the two input neurons which respond to the occurrence of *there* and *on* are shown, connected to one of the hidden layer neurons which in turn stimulates the two outer layer neurons.

These artificial neurons differ from human neurons in at least three important respects:

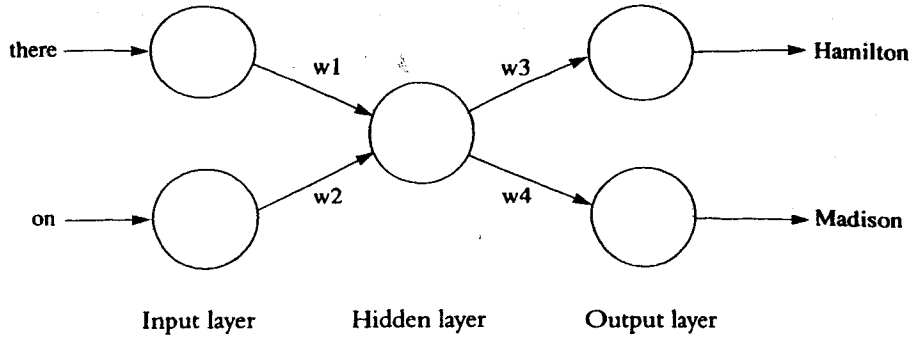

Figure 5.2 A section of the Federalist neural network

1. In the artificial network, the degree of stimulation afforded by one neuron to another to which it is connected depends on a weight which is continually updated during the training phase.
2. The weights in Figure 5.2 are labelled $w1$ to $w4$. The human neuron is always in one of two states. It either fires if the sum of its inputs are greater than a certain threshold, or remains inactive if the sum of its inputs is below that threshold. In this artificial network, however, the neuron becomes active to an extent which gradually increases as the sum of its inputs increases.
3. The human neuron normally has only one output channel or 'axon', but the artificial neurons can stimulate any number of other neurons.

Taking (1) and (2) into account, the degree of stimulation produced by an artificial neuron depends on the strength of both its inputs and its weight. The Federalist network was trained by a process called the conjugate gradient method. Starting with random weights, the network was fed data pertaining to the occurrence of the marker words in a text of known authorship, and the network would select either Hamilton or Madison as the most probable author. According to whether the network was correct or not, the weights were adjusted, then the network was given the marker-word data from another text. This process continued until the network had been 'shown' 100 documents. The weights were then frozen, and the testing phase began. The 12 disputed papers were presented to the network in turn, and in each case they were classified as being by Madison.

Holmes and Forsyth (1995) also studied the Federalist papers, using genetic algorithms which, like neural networks, are inspired by biology. Within the genetic algorithm, a set of rules, in this case a set of discriminants for distinguishing the work of different authors, is likened to a set of biological genes. The system starts with a random set of rules. During the training phase of the algorithm, the system attempts to determine the authorship of the non-