50. Corpus linguistics and stylometry

1. Introduction
2. Lexical discriminators
3. Statistical measures
4. Automatic selection of discriminators
5. Conclusion
6. Acknowledgments
7. Literature

1. Introduction

Computational Stylometry is an attempt to capture the essence of the **style** of a particular author by reference to a variety of quantitative criteria, usually lexical, called **discriminators.** The most common application of computational stylometry has been author identification in cases of disputed authorship. The underlying assumption in studies of authorship is that although authors may consciously vary their own style, there will always be the subconscious consistent use of certain stylistic features throughout their work (Holmes 1997). However, there is no clear and indisputable evidence that such features exist (McEnery/Oakes 2000). Discriminators used in quantitative stylometric studies must occur frequently in the texts, and represent traits that can be expressed numerically. The identification of features such as Semitisms in the New Testament is a subjective affair, and hence these would not be good discriminators (Morton 1978). Discriminators which have been used in studies of stylometry include:

1. **Word and sentence length.** In perhaps the earliest study, in 1887, Mendenhall showed that the modal word length was 2 in the writings of J. S. Mill, but 3 for *Oliver Twist* (Kenny 1982). These measures tend to work less well for studies of disputed authorship, since they are more under the conscious control of the author. They work better as discriminators of genre or register, as a comparison of the texts used in different newspapers will show.
2. **Vocabulary studies**: The choice and frequency of words, and measures of vocabulary richness such as Yule's K measure (Yule 1944, 57). The overwhelming majority of corpus-based stylometry studies use lexical (single word) discriminators, so we will look at these more closely in section 2.
3. **Fragments of words,** such as bigrams, or pairs of adjacent characters (Kjell 1994). There is a whole literature concerned with discovering automatically which word fragments might act as the best discriminators, and we will explore this in section 4.
4. **Words commencing with an initial vowel** (Hilton/Holmes 1993).
5. **Collocations of words.** (Hoover 2002, 2003) used pairs of words in his studies of disputed authorship, both contiguous sequences of two words, and collocations, which he defined as "any two words that occur repeatedly within a specified distance from each other, counted in words". In some cases he found that word pairs gave better results than single words, but in others not.
6. **Positions of words within sentences.** In Michaelson/Morton's chronology of Isocrates (1976), the frequency of use of *gar* (meaning *for*) as the second or third word showed

a negative correlation with time. Milić (1966) describes "unconscious ordering in the prose of Swift".

7. **Syntactic analysis**. Such analyses require the preparation of syntactically annotated corpora. For example, Antosch (1969) showed that a high adjective to verb ratio was found in folk tales, but a much lower ratio was found in scientific texts. Baayen/Van Halteren/Tweedie (1996) studied the frequencies of use of phrase structure rewrite rules in annotated corpora to distinguish between crime fiction authors, and Santini (2004) extracted syntactic features for genre classification.
8. **Pause patterns in Shakespeare's verse** (Jackson 2002).

The ideal situation for authorship studies is when there are large amounts of undisputed text, and few contenders for the authorship of the disputed text(s). The **experimental methodology** for determining which of two authors are the more likely to have written a newly-discovered or disputed text is as follows: Build corpora A and B, containing texts undisputedly written by authors A and B respectively, and then build corpus C consisting of works of disputed authorship, but probably written by either A or B. Then select a set of **discriminators** and an appropriate **statistical measure.** When these have been shown to discriminate effectively between A and B, try them on corpus C, to see whether the works in corpus C more closely resemble those in corpus A or those in corpus B. This method of distinguishing between authors also works for other differences in authorial style such as gender. Corpus A would contain texts known to have been written by female authors, Corpus B would contain texts definitely written by male authors, and a suitable choice of discriminators and a statistical test would show whether corpus C, containing text(s) by an unknown author, has characteristics more typical of male or female authorship.

A number of modern (post 1945) developments have enabled great advances in stylometry. Modern statistical techniques enable us to look for significant differences between the data sets as opposed to chance variation, such as the t-test (Binongo/Smith 1999) and the z-score (Burrows 2002). Modern sampling techniques mean that we have no need to examine the entire works of a given author before making inferences about that author's style. Thirdly, the advent of computers has enabled fast and accurate calculations, and storage of large text corpora. To date, there are no commercial computer packages for stylometry. Many studies have developed analyses of texts by hand, or using simple frequency counting programs written in languages such as Perl (which has been designed especially for text handling), they have extracted the relevant data and processed it using statistical software packages (such as SPSS or MATLAB, see article 36) or manual statistical analysis.

The major difficulty with identifying the style of an individual author is that an individual style can be masked by a number of related issues, such as the following:

1. Heterogeneity of authorship over **time.** A number of authors have been shown to vary in their stylistic traits over time. The earliest such study producing a chronology of texts, by Yardi (1946) showed that certain features of Shakespeare's writing varied as he got older. In this article we will look more closely at how discriminators were chosen to distinguish between the younger and the older Yeats (Forsyth 1999).
2. Authorship and **genre.** Genre differences have been found to be more pronounced than author differences. Thus Baayen/Van Halteren/Tweedie (1996) ensured that in order to discriminate between a pair of authors, the comparison texts were all in the genre of crime fiction.

3. Authorship and **gender.** Rayson/Leech/Hodges (1997) showed differences in the vocabulary used by men and women in the spoken component of the British National Corpus, and Koppel/Argamon/Shimoni (2002) were able to automatically categorise texts by author gender.
4. Variation **within a single author**. A number of very versatile authors, such as Jane Austen (DeForest/Johnson 2001) and Oliver Goldsmith (Dixon/Mannion 1993), show consistent variations between the characters of their novels. A related problem is that traits of a school of writers (such as the school of Anglo-Irish writers of which Goldsmith was a member) can overshadow any personal tendencies. According to Laan, "[i]deally, a study of the differences in style within the works of an author should precede an attribution or chronology study concerning that same author" (Laan 1995, 271).
5. In the field of Information retrieval, which is concerned with searching on the internet, the text categorisation literature is more interested in **categorisation by topic** than by writing style. Categorisation by topic is typically based on medium frequency keywords that reflect the content of the document. However, categorisation of texts by author style uses precisely those features (such as high frequency function words) that are independent of content (Koppel/Argamon/Shimoni 2002).

Despite these competing influences, there have been many successful studies of authorship attribution.

2. Lexical discriminators

Qualitative studies of literary style have focused on the **hapax legomena,** the words which appear only once in the entire text. These tend to be obscure, out-of-date or technical terms, or convey delicate shades of meaning, and thus reflect the background and experience of the author. They form the largest group of words in the vocabulary of a text. For example, in a 100,000 word sample of the British National Corpus, about 9500 are hapax legomena, while only about 1500 words come up twice, about 1000 three times, about 600 four times, and about 300 five times. The problem with the hapax legomena as discriminators is that their individual low rate of occurrence makes them difficult to handle statistically. Statistical tests such as the chi-squared test require as a very minimum five occurrences of a discriminator in at least one of the corpora. Thus **quantitative studies** of literary style which use lexical discriminators must make use of the words which appear frequently in the texts.

2.1. Vocabulary richness

Many quantitative studies rely on the concept of **vocabulary richness**. A text has low vocabulary richness if the same limited vocabulary is repeated over and over again, while it has high vocabulary richness if new words continually appear. In the following discussion of **measures of vocabulary richness,** we make use of the following notation:

(a) tokens N = length of text in words
(b) types V = number of different words in the text

(c) hapax legomena $V1$ = number of words occurring just once in the text
(d) dislegomena $V2$ = number of words occurring exactly twice in the text
(e) Vi = number of words occurring exactly i times

The **type / token** ratio depends on the length of the text (being generally less for longer texts), but is a useful measure of vocabulary richness when the comparison texts are of equal length. **Honoré's measure R** (Honoré 1979) depends on the hapax legomena:

$$R = 100 \log N / (1 - (V1 / V))$$

Sichel's measure S (Sichel 1975) depends on the dislegomena, and is relatively constant with respect to N:

$$S = V2 / V$$

Brunet's measure W (Brunet 1978) is:

$$W = N^{V-a},$$

where a is a constant (usually 0.17). W was found to be relatively unaffected by text length and to be author specific (Brunet 1978). **Yule's characteristic K** depends on words of all frequencies:

$$K = 10{,}000 * (M - N) / N^2, \text{ where } M = \Sigma\, i^2 \,.Vi.$$

Yule (1944) used his characteristic K to determine whether *De Imiatione Christi* was more likely to have been written by Kempis or Gerson. The results (see Table 50.1) show that the vocabulary richness of *De Imitatione Christi* is much closer to that of the works by Kempis than that of those by Gerson, and hence Kempis is the more likely author.

Tab. 50.1: Yule's Characteristic K for De Imitatione Christi

De Imitatione Christi	K = 84.2
Works definitely by Kempis	K = 59.7
Works definitely by Gerson	K = 35.9

Holmes (1992) performed a stylometric analysis of **Mormon Scripture** and related texts, which used five measures of vocabulary richness: Honoré's R, Yule's K, Sichel's S and two associated parameters called α and θ. His results, averaged over all five measures, are shown in Table 50.2, where all the texts apart from the personal writings of Joseph Smith (the movement's founder) and the King James Bible are samples of Mormon Scripture.

These scores were used to **cluster** the texts, producing a pictorial representation called a **dendrogram** (because it looks like a tree) where texts similar in vocabulary richness would appear close together, and those dissimilar in vocabulary richness would appear far apart. This involved using the raw values to produce a **similarity matrix,** which stored the similarity between each pair of text samples, using the formula

$$1 - ((Xr - Xs) / \text{range})^2$$

Tab. 50.2: Yule's Characteristic K for Mormon scripture and related texts

J. Smith − personal 1 (J1)	K = 57.7	Mormon 3 (M3)	K = 119.2
J. Smith − personal 2 (J2)	K = 82.1	Mormon 4 (M4)	K = 168.9
J. Smith − personal 3 (J3)	K = 78.6	Mormon 5 (M5)	K = 125.5
Nephi 1 (N1)	K = 145.2	Alma 1 (A1)	K = 149.0
Nephi 2 (N2)	K = 155.2	Alma 2 (A2)	K = 150.6
Nephi 3 (N3)	K = 150.5	Doctrine 1 (D1)	K = 126.9
Jacob (JB)	K = 134.3	Doctrine 2 (D2)	K = 91.6
Lehi (LI)	K = 109.4	Doctrine 3 (D3)	K = 98.9
Moroni 1 (R1)	K = 131.5	Isaiah 1 − King James (I1)	K = 81.3
Moroni 2 (R2)	K = 115.7	Isaiah 2 − King James (I2)	K = 114.2
Mormon 1 (M1)	K = 183.8	Isaiah 3 − King James (I3)	K = 90.9
Mormon 2 (M2)	K = 132.7	Book of Abraham (AB)	K = 146.4

For example, using only the data for Yule's characteristic *K*, if text *R* is Jacob, and text *S* is Lehi, $Xr - Xs = 134.3 - 109.4 = 24.9$. The range is the difference between Yule's *K* characteristic for the text with the richest vocabulary and the text which was most sparse in vocabulary, which is $183.8 - 57.7 = 126.1$. According to the formula, the similarity between Jacob and Lehi is $1 - (24.9 / 126.1)^2 = 0.96$. Two texts identical in vocabulary richness would have a similarity of 1, while the pair of texts most dissimilar in vocabulary richness has a similarity of 0. The similarity scores were averaged over all five measures. From the similarity matrix, the dendrogram is produced step by step as follows. First the two most similar texts are joined together to form a cluster. Then the next most similar pair of texts is joined together, but if the similarity between the newly formed cluster and a single text is greater than the similarity between any two single texts, then the cluster is joined to that single text to form a larger cluster. The process

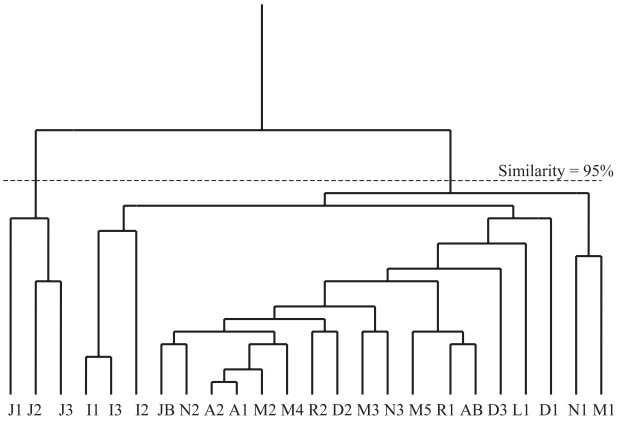

Fig. 50.1: Dendrogram for the total data set

continues, with the most similar clusters or single texts being joined at each stage, until all the texts belong to a single cluster.

In Holmes' study, the dendrogram (which is reproduced in Figure 50.1) showed that distinct subtrees were found for each of Joseph Smith's personal writings, Isaiah, and the Mormon prophets. However, variation within each Mormon prophet's writings was greater than the variation between prophets. Commenting on this, Holmes stated that he has found no evidence of any multiple authorship in the Book of Mormon, though it was purportedly written by a variety of different prophets at different times. He suggested Joseph Smith's personal writings differed from the Book of Mormon in that he had adopted the style of a "prophetic voice" when writing the Book of Mormon (Holmes 1991). In another study, Pollatschek/Radday (1985) used vocabulary richness to examine the authorship of Genesis. Weaknesses of methods based on vocabulary richness are discussed by Hoover (2004).

2.2. Individual words

Many studies have looked at authors' use of **function** words, which are high-frequency closed-class vocabulary items. A standard list of such words is Taylor's list of ten function words (T10): *but, by, for, no, not, so, that, the, to, with.* Merriam/Matthews (1993) used ratios of these: *no/T10, (of x and)/of, so/T10, (the x and)/the, with/T10* in a study which showed that an anonymous play, Edward III, was more likely to have been written by Shakespeare than Marlowe. It is believed by many that function words are less under the conscious control of the author compared with rare words. Mosteller/Wallace, in their study of the Federalist Papers (described in more detail in section 3.1.), used 30 individual words which were used much more often by either Hamilton or Madison (the two possible authors of the disputed papers), such as *direction, innovation(s), language, vigor(ous), kind, matter(s), particularly, probability, work(s).*

DeForest/Johnson (2001) write that the proportion of **Latinate** words to words of Germanic origin in an English text can be used as a stylometric measure. This technique requires the compilation of a large dictionary of words and their origins: DeForest/Johnson classified all 13,809 unique words in Jane Austen's writings. The characters in the novels were found to vary in the proportion of Latinate words they used, with high proportions of Latinate words being indicative of high social class, formality, insincerity and euphemism, self-control (as opposed to emotion), men's speech (since education was the preserve of men in the 18th century) and stateliness as opposed to squalor (for example, contrasting Mansfield Park with the less loved house in Portsmouth). This study is also discussed in article 6.

Collins et al. (2004) used a corpus tagged for **rhetorical language choices**, and found that the writings of Hamilton and Madison in the Federalist Papers differed most in their use of "Think positive" and "Think negative" (used more by Hamilton) and "Past events" (used more by Madison).

3. Statistical measures

Having discussed the choice of discriminators, we must now consider the choice of statistical measure. Standard statistical tests, such as the chi-squared test, the z-score or the

t-test, help us to decide whether differences found in the use of discriminators by different authors genuinely show an underlying pattern, or whether they have resulted from relatively small chance variations in the data. Other statistical tests, particularly the multivariate methods such as clustering (Holmes 1992), Correspondence analysis (Mealand 1995) or Principal Components Analysis (PCA) provide pictorial representations, ideally much clearer than the raw data, from which subjective judgements can be made. Related techniques are described in articles 38 and 40. Although not statistical tests as such, we will discuss neural networks and genetic algorithms, used in artificial intelligence, but originally inspired by biology. These models perform the role of statistical tests in stylometry, in that they take in raw data about the frequency of discriminators in a text, and produce as output a decision about the authorship of that text.

3.1. Bayesian probability

Bayesian probability was used by Mosteller/Wallace (1964) to examine a case of disputed authorship in the Federalist Papers. The Federalist Papers were published under the pseudonym "Publius" in 1787−1788 to persuade the people of New York to accept the new American constitution. It is undisputed that Jay wrote 5 of the essays, Hamilton wrote 43, and Madison wrote 14. However, 12 of the essays are disputed. There is some historical evidence that Madison was the author of these twelve essays, but Hamilton, on the night before he was killed in a duel, left a list of the essays and their authors at a friend's house. On this list, the 12 disputed essays were attributed to Hamilton. Mosteller/Wallace found that the styles of Hamilton and Madison varied in the frequency of use of certain words. For example, *enough* was found in 14 papers by Hamilton, but none by Madison; *whilst* was found in no papers by Hamilton, but in 13 by Madison. 30 such discriminating terms were found. They proceeded as follows (Francis 1966):

If we know the average number of times a word appears in a text of fixed length, we can find the proportion of text sections of that length which have none, one, two, etc. occurrences of that word using the **Poisson distribution**:

$$P_n = \frac{\lambda^n . e^{-\lambda}}{n!}$$

e to the power $-\lambda$ is the $exp(-\lambda)$ on a calculator. λ is the average number of times the word occurs per section of text, while Pn is the proportion of text sections which have n occurrences of the word. Suppose the average rate of use of *also* is 0.5 words per 2000 for Hamilton and 1.0 words per 2000 for Madison, and suppose, too, that the word appears four times in a disputed paper of length 2000 words.

The probability of *also* occurring exactly four times in text by Hamilton is denoted by P_4:

$$P_4 = \frac{\lambda^n . e^{-\lambda}}{4!} = \frac{0.5^4 . e^{-0.5}}{24} = 0.00158$$

The corresponding calculation for Madison is:

$$P_4 = \frac{1 . e^{-1}}{24} = 0.0153$$

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. This evidence is combined with other evidence, such as historical evidence giving initial odds of 3 to 1, and other words with high discriminating power, such as *an* which occurs 7 times in the unknown document, thus favouring Madison with odds of about 8 to 3. These three pieces of evidence can be combined using Bayes' theorem, giving odds

$$\frac{3}{1} x \frac{10}{1} x \frac{8}{3} = \frac{80}{1} \text{ in favour of Madison.}$$

Mosteller/Wallace judged all 12 disputed papers to have been written by Madison.

3.2. Univariate analyses

All the common univariate analyses (which get their name because only the variation in the use of one discriminator between authors is considered) such as the t-test are covered in standard statistics text books, such as the one by Woods/Fletcher/Hughes (1986). However, we will include the **z-score** here as an example, with data derived from a study by Burrows (2002). To understand the z-score, we must be familiar with the notions of the mean and the standard deviation. The mean is the average of all values in a data set, found by adding up all of the values and then dividing by the number of values. The standard deviation is a measure which takes into account the distance of every data item from the mean. If all the values in the data set are exactly equal to the mean, then the standard deviation is 0, otherwise, if they vary from each other, standard deviation will be more than 0.

Using Burrows' (2002) data, a sample of 25 Restoration writers shows that the mean occurrence of the word *the* is 4.719%, and the standard deviation is 0.63%. In one of these text samples (actually taken from Milton's *Paradise Lost*), the occurrence of *the* is 4.242%. We can now calculate the measure called the z-score, as follows:

$$z = \frac{x - \bar{x}}{s} = \frac{4.719 - 4.242}{0.63} = 0.757$$

This value can be looked up in a normal distribution table (there is one at the back of most statistics textbooks), to show that we would expect a sample which belongs to the collection of Restoration writers to have a z-score of 0.757 or more about 45% of the time. Only if the resulting z-score would be expected to occur in the collection just 5% of the time or less, would we suspect that the Milton sample did not really belong to this collection. Burrows extends the z-score idea to produce the delta score, designed as a measure capable of distinguishing the most likely candidate from a large group, while traditional studies discriminate only among a small number of possible authors.

3.3. Cumulative sum charts

A. Q. Morton believed that the rate of occurrence of a (stylistic) 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. This is the rationale behind Morton and Michaelson's (1990) controversial cusum (cumulative sum) technique which has been accepted by several courts of law in cases revolving around allegedly forged confessions, such as the successful appeal against a robbery conviction by Tommy McCrossen in 1991. However, in another case where Bob Maynard and Reg Dudley were accused of two London gangland murders, Morton's evidence in their favour was successfully rebutted at the Court of Appeal (Campbell 1992, 1997).

Two plots are drawn on the same graph − one corresponding to the lengths of the sentences in the text under study, and one corresponding to the occurrence of some other linguistic feature, such as the number of words with initial vowels. The cusum values s_i to be plotted for each sentence i are given by the formula

$$s_i = \sum_{r=1}^{i} (x_r - \bar{x})$$

where r refers to the individual sentences from the start of the text up to and including sentence i, x_r is either the sentence length or the number of times the chosen linguistic feature is found in sentence r, and $\bar{x}$ is either the average (mean) sentence length for that text, or the average number of times the linguistic feature is found per sentence in that text. On the cusum plot, s_i is plotted on the vertical axis, while i is plotted on the horizontal axis. The theory is that if a work was written by a single author, the ratio of occurrences of the linguistic feature to sentence length will be relatively constant, and thus the two lines drawn on the cusum chart, if suitably scaled on the vertical axis, will follow each other almost exactly. On the other hand, if the early part of text was written by one author, and the later part of the text by another, the two lines on the cusum chart will start to diverge at the point where the authorship changes. In the example shown in Figure 50.2, from Hilton/Holmes (1993), there is a divergence in the lines near to sentence 25, the point at which a sample of *Northanger Abbey* ends and a sample of *The Great Gatsby* begins.

On two occasions rebuttal evidence against the cusum technique has been prepared for the crown by David Canter (Campbell 1992). Canter (1992) showed that the technique was not reliable, whether one judges by a subjective comparison of the two plots on the cusum chart, or attempts to quantify the correlation between the two lines using Spearman's rank correlation coefficient. Holmes/Tweedie (1995) also give a critique of the cusum controversy. To overcome the subjectivity of the technique, Hilton/Holmes proposed the weighted cumulative sums technique, where the two lines of the cusum chart are combined into a single line. They use a version of the t-test to see if there is a significant difference between an earlier and a later portion of the line. Although this puts the technique on surer statistical foundations, Hilton/Holmes found that the weighted cusum technique performed only marginally better that the cusum test, and that neither technique gave consistently reliable results. They concluded that authors are not as consistent in their selection of linguistic features as would be required for cusum techniques to determine authorship correctly.

Barr (1997) found that cusum graphs were useful for examining scale differences between authors. As an example of scale differences, one author might have a typical

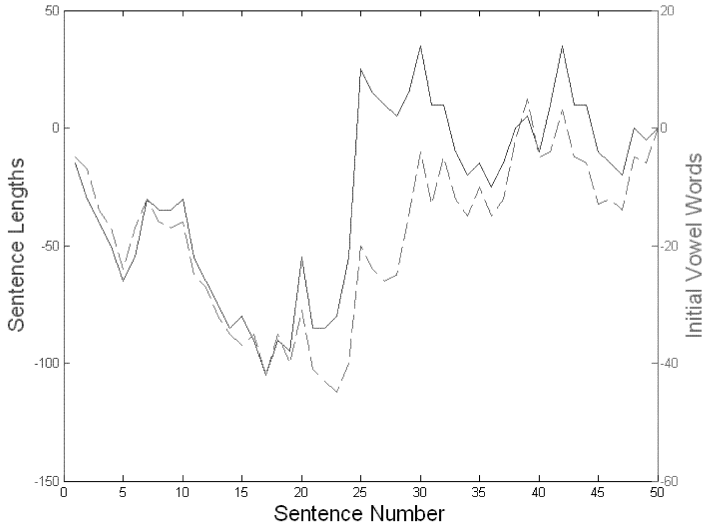

Fig. 50.2: Cusum Plot for Northanger Abbey and The Great Gatsby. There is a significant discrepancy in the plot near sentence twenty-five, where the samples are concatenated. The solid line is the plot of sentence lengths, while the dotted line is the plot of words with initial vowels

opening pattern of long sentences, a middle pattern of sentences of mixed length, and a closing pattern of short sentences. These proportions might be retained in other works by the same author, even if these other works are of varying length.

3.4. Multivariate analysis

Multivariate analyses (see also articles 38 and 40) are so-called because they take into account a large number of variables or discriminators. If each discriminator can be thought of as a dimension, multivariate analyses attempt a **reduction in dimensionality.** If several discriminators tend to vary together, i. e. in texts where one is frequently found, the others are often found also, they are combined into a single dimension (called a factor or principal component, according to the type of analysis being performed), thereby achieving this reduction in dimensionality. For example, we could start with four dimensions (though a real study might use many more), being the whole-word discriminators *he, her, him, she*. These words might be distributed across five texts as shown in Table 50.3.

Here we see that the discriminators *he* and *him* tend to be found in the same texts as each other, and so do *her* and *she*. Thus we can reduce the original four dimensions to just two (one for *he*/*him*, and one for *her*/*she*). Here there are only two dimensions left, but in a typical study, the two most important dimensions (in terms of accounting for the variation in the data) will account for about half the total variation, and the less important dimensions are disregarded. The advantage of cutting down to two dimensions is that they can become the two axes of a graph, on which all the texts can be plotted. Text 1 contains 22 words in the *he*/*him* dimension, and 2 in the *her*/*she* dimen-

Tab. 50.3: Candidate data set for dimensionality reduction

	He	Her	Him	She
Text 1	10	2	12	0
Text 2	11	1	15	3
Text 3	9	1	10	1
Text 4	0	14	0	14
Text 5	1	12	2	11

sion. Thus it can be plotted at point (22,2) on the graph. (In real life, the correspondence between the raw data and the position on the dimension axis is not so exact). Once all the texts have been positioned on the graph, they will hopefully appear in clusters, where all the texts written by one author will be positioned close together, clearly set apart from the texts by other authors.

In this section we will consider a study which used **Principal Components Analysis** (PCA), although Factor Analysis and Correspondence Analysis also follow this dimensionality reduction and two-dimensional plot paradigm. Burrows (2002) found that a problem of PCA was that if one of the texts was added to or removed from the analysis, the whole pattern can alter so that we are no longer able to make strict comparisons between graph and graph. However, "in experienced hands, such methods yield excellent results" (ibid., 269). It is important to perform some sort of cross-validation, to ensure that the main clusters remain, when each single text is removed in turn.

Fifty years after the American Civil war, General George Pickett's widow, LaSalle Corbell Pickett, published letters purportedly written by her husband, many of them written during his active service in the war. Historians were divided as to their authenticity, so Holmes/Gordon/Wilson (2001) used PCA for a stylometric investigation into the Pickett letters. Starting with the 60 most common words in the texts as discriminators, the analysis produced two principal components. Various text samples were plotted on the resulting two-dimensional plot for comparison, as shown in Figure 50.3.

(a) auto, LaSalle Pickett's autobiography
(b) ltr, LaSalle Pickett's personal letters
(c) gp, George Pickett's personal pre-war and post-war letters
(d) hs, the disputed letters
(e) gw, George Pickett's war reports
(f) i, Inman papers, genuine handwritten letters by George Pickett
(g) har, Walter Harrison's book "Pickett's men" (one theory being that the letters were plagiarised from this book).

The investigation strongly suggests that LaSalle Pickett composed the published letters herself. All the seven sources listed above are internally consistent, forming clusters, and the samples of LaSalle Pickett's autobiography fall suspiciously close to the disputed letters. Tweedie/Holmes/Corns (1998) used PCA to examine the provenance of *De Doctrina Christiana,* traditionally attributed to John Milton. Mealand (1995) carried out a Correspondence Analysis of *Luke* confirming modern theological opinion on the

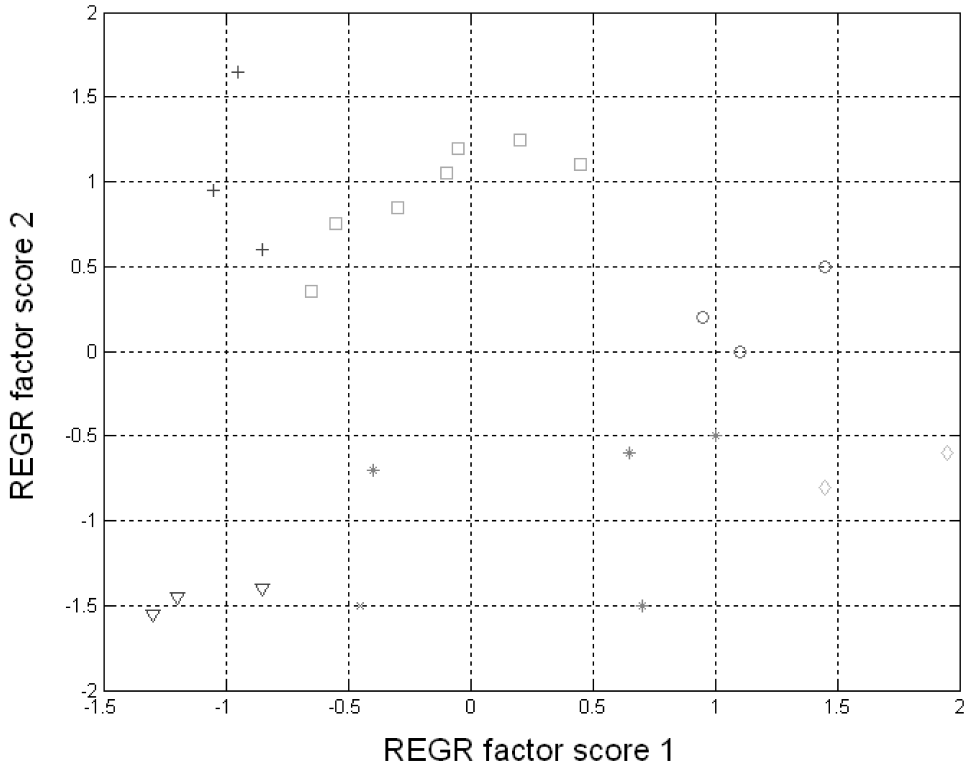

Fig. 50.3: Principal components plot: attribution. + = autobiography, o = LaSalle letters, * = George personal, x = George war, square = "Heart", diamond = Inman papers, triangle = Harrison.

sources of that gospel. Factor analysis has been successfully used for the study of register variation in English (Biber 1995). Clustering (see section 2.1.) is also a form of multivariate analysis.

3.5. Neural networks

A typical neural network architecture, called a multi-layer perceptron, is exemplified by Figure 50.4, taken from Matthews/Merriam (1993). It consists of three layers of nodes: the input layer has five nodes, one for each discriminator used in the study, there is a middle, "hidden" layer, and an output layer of two nodes (one for each possible author of a text). For each text, the input nodes are activated in proportion to how often that discriminator is found in the text. This activation is passed on to the middle layer, according to the strengths of connection (or **weights**) between the nodes of those two layers. In turn, the activation is passed on to the outer layer − one node will be activated more if the network "thinks" that author 1 wrote the text, while the other will be activated more if it seems that author 2 wrote the text. This will only work if the weights

are correct. They are initially random, but in a prior "training" session, the weights are gradually updated in response to the frequency of discriminators and the identity of authors of known texts, until the network is right every time. The unknown texts (the "test data") are then presented to the network one by one, and in each case the network gives an "opinion" (author 1 or author 2) as to who wrote that text.

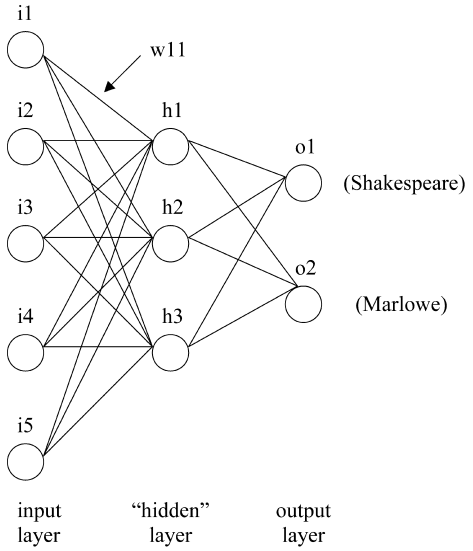

Fig. 50.4: A stylometric multi-layer perceptron

Matthews/Merriam (1993) created a neural network to distinguish Shakespeare and Fletcher. Hoorn et al. (1999) produced networks to distinguish three Dutch poets, Bloem, Slauerhoff and Lucebert. Waugh/Adams/Tweedie (2000) discuss how to minimise the number of nodes in the hidden layer for a stylometric study. Kjell's (1994) study is interesting, because his network has 26×26 input nodes, one for each possible bigram or pair of adjacent characters. This means that there is no need to manually create a set of discriminators on an "author pair" basis. After training, bigrams which have not proved to be useful discriminators have been given zero weights, and thus by process of elimination, the network has "learned" a set of good discriminators.

3.6. Genetic algorithms

Another model based on biological analogy is the genetic algorithm, based on the Darwinian idea of Natural Selection. Forsyth's PC/BEAGLE system, originally devised for weather forecasting, was applied to stylometry in a new study on the Federalist Papers (Holmes/Forsyth 1995). It consisted of a set of **rules**, made up of operators, variables and constants, in the form "IF such-and-such is true, THEN the essay was written by Hamilton". For example $(KIND < 0.00002)$ & $(TO < 35.205) \rightarrow$ *Hamilton* means that if the word *kind* appears less than 0.00002 times per thousand words, and also the word

to appears less than 35.205 times per thousand, then the article must have been written by Hamilton. Another example would be *(UPON − BOTH) < WHILST → Hamilton.* The word variables are drawn from Mosteller/Wallace's (1964) 30 marker words. PC/BEAGLE's learning algorithm is given below:

1. Create an initial population of candidate rules at random (all of which must be syntactically valid, but most of which will be semantically meaningless).
2. Evaluate each rule on every training example (texts of known authorship) and compute a fitness score, based on how often the rule predicts the author correctly, with a penalty for rule length to encourage brevity.
3. Rank the rules in descending order of merit and remove the bottom half.
4. Replace discarded rules by crossing a pair of randomly selected survivors to produce "offspring". This "mating" is achieved by picking out a random sub-expression from each of two surviving rules and tying them together with a randomly chosen connective. Possible descendents of the two sample rules given above could be *(UPON − BOTH) < 35.205 → Hamilton*, and *(KIND < 0.00002) & (TO < WHILST) → Hamilton.*
5. Mutate a small number of rules picked at random (excluding the best rule), by changing one component, and apply a tidying procedure to reduce redundancy. If the termination conditions are met (e. g. a new generation of rules has not shown any improvement over the last generation), training is complete. Otherwise, return to stage (2).

Once training is complete, the set of rules can be applied to cases that may not have been seen before, in order to determine their authorship. As with Kjell's neural network, the genetic algorithm, apart from being restricted to the 30 marker words, is free to choose its own set of discriminators.

4. Automatic selection of discriminators

There are two main advantages of selecting discriminators automatically, namely that to do this manually is time consuming, and also that it results in sets of discriminators that work on one author pair but not necessarily on others. The process is necessarily subjective, and thus each stylometrist might have a "tool-kit" of favourite marker types, leading them to overlook the vast majority of those discriminators that might be used (Holmes/Forsyth 1995). A fourth reason is given by Burrows (2002, 268):

> "a wealth of variables, many of which may be weak discriminators, almost always offer more tenable results than a smaller number of strong ones. Strong features, perhaps, are easily recognised and modified by an author and just as easily adopted by disciples and imitators. At all events, a distinctive stylistic signature is usually made up of many tiny strokes."

Koppel/Argamon/Shimoni (2002) explore the possibility of automatically classifying formal written texts by author **gender.** They start with 1081 features, chosen solely for relative topic independence: 405 function words, the 500 commonest part-of-speech trigrams, the 100 commonest bigrams, and all 76 single part of speech categories. Using their "Winnow" learning method, they were able to iteratively discard the features which

were not good discrimators. The last features to die off for fiction texts were *a, the, as* for male authors; *she, for, with, not* for female authors. When training on non-fiction they found the last features to disappear were *that, one* for male authors, and *for, with, not, and, in* for female authors. For parts of speech the male indicators were determiners, numbers and modifiers, while the most effective female discriminators were negation, pronouns and some prepositions. Best results for determining the gender of the authors of new texts was obtained when 64 to 128 features were retained, which gave an accuracy of 84% for non-fiction, and 80% for fiction.

As we saw in sections 3.5. and 3.6., neural networks and genetic algorithms are able to learn their own sets of discriminators. In sections 4.1. and 4.2. respectively, we will look at machine learning techniques for selecting whole word and character substring discriminators (for machine learning techniques see also article 39).

4.1. Whole words elimination

Five methods are included in the text classification study by Yang/Pedersen (1997), namely document frequency (DF), information gain (IG), mutual information (MI), a chi-squared statistic (CHI) and term strength (TS). In each case, every word in the texts to be classified is initially a potential discriminator, but the majority of these are discarded by giving each one a numeric score. If, for example, we wanted to reduce the number of discriminators to 100, then only the 100 highest scoring words would be retained. Only the first method, **document frequency**, is unsuitable for stylometry, since it selects the mid-frequency terms that are more useful for detecting the topic rather than the writing style.

Information gain measures the amount of information, measured in bits, obtained for category prediction by knowing the presence or absence of that term in a document. To calculate **mutual information**, consider the two-way contingency table of a word w and author a, where A is the number of times word w is used by the first author, B is the number of times word w is used by the other author, C is the number of times the first author used any term other than w, and D is the number of times the other author uses any word other than w. N is the total number of word tokens in the entire data set. $MI(w,a)$, the mutual information between a word and an author, is estimated using

$$MI(w,a) \approx \log \frac{AN}{(A+C)(A+D)}$$

$MI(w,a) = 0$ if the word and the author are independent of each other, but has a positive value if the use of the word suggests that author, and a negative value if the author consistently avoids the use of that word. The problem with this measure is that MI favours rare words. Using the same contingency table, we can work out the **chi-squared** statistic:

$$X^2(w,a) = \frac{N(AD-CB)_2}{(A+C)(B+D)(A+B)(C+D)}$$

Chi-squared = 0 if the term and the category are independent. The chi-square statistic is also known not to be reliable for low-frequency terms. Calculation of Yang/Pederson's fifth measure, **term strength**, is more complex.

Binongo/Smith (1999) describe the use of the **t-test** (independent samples, i. e. non-matched pairs) to find words that discriminate between ten blocks of text by Shakespeare and five by Wilkins. The best discriminants (those producing the highest t-scores) were in order: *then, the, by, a/an/awhile, on, no, most, to/into, there* and *for/forever.* Classes of words not considered were verbs (due to their various inflected forms), nouns and personal pronouns (which are often context-dependent) and rare words. For example, using data provided by Binongo/Smith, the occurrence of the word *the* in five samples of *Cymbeline* (Shakespeare) is 166, 163, 165, 177 and 174. The same word is found in four samples of Wilkins' *The Miseries of Enforced Marriage*, 97, 112, 138 and 112 times. Every sample consisted of 5000 words. The mean and standard deviation are 169 and 6.12 for *Cymbeline*, and 114.75 and 17.04 for *Miseries.* These two standard deviations are combined to produce a common standard deviation, *s*:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n^2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(5 - 1).169^2 + (4 - 1).114.75^2}{5 + 4 - 2}} = 12.44$$

This value is then used in the calculation of *t*, as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}} = \frac{169 - 114.75}{\sqrt{\frac{12.44^2}{5} + \frac{12.44^2}{4}}} = 6.501$$

The corresponding calculation for *and* yields a *t* value of only 2.097, so the occurrence rate of *the* is a better discriminator than the occurrence rate of *and* for *Cymbeline* and *Miseries*.

4.2. Monte-Carlo Feature Finding

Assigning a date to a text is the main task of **stylochronometry.** In this section we will look at Forsyth's (1999) study which compared the writing of the older (post-1915) and the younger Yeats. He aimed to first develop a stylochronometric technique for an author whose chronology is well established, so that this technique could later be used for authors whose dating is less well documented. Yeats himself insisted that his language changed as he grew older, and most readers would agree, but what exactly were these changes? To find out, Forsyth used an algorithm, called **Monte-Carlo Feature Finding (MCFF)**. Initially, all character sequences of eight characters or less, found in at least one of the poems chosen for the training set, are regarded as potential discriminators. Since there are so many of these, a random sample of just 4096 was taken. Each of these substrings were ranked according to their **distinctiveness**, as measured by the **chi-squared test** (see section 4.1.). The training data was divided into two portions: 72 poems repre-

Tab. 50.4: Top six discriminators for the younger and older Yeats

Rank	Substring	Chi-squared	YY-count	OY-count
1	“what”	35.1	30	100
2	“can”	34.3	21	82
3	“s, an”	25.4	63	19
4	“whi”	25.4	67	21
5	“with”	22.3	139	74
6	“?”	21.9	30	83

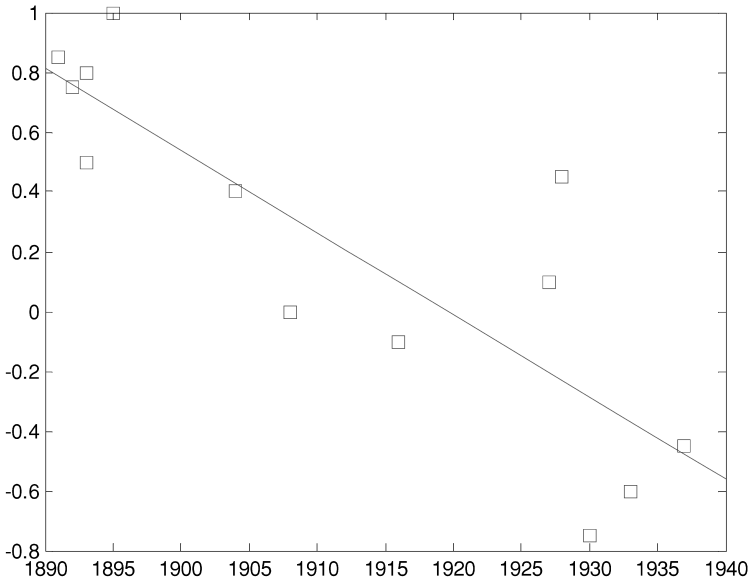

Fig. 50.5: Plot of Youthful Yeatsian Index against date

senting the younger Yeats, and 70 poems representing the older Yeats. 88 distinctive substrings were identified, and the most distinctive of all are tabulated in Table 50.4.

The columns labelled YY-count and OY-count show how often each substring was found in the younger and older Yeats samples respectively. The counts of all 88 retained substrings were then found in thirteen other poems written between 1891 and 1931, and a plot was drawn of “youthful Yeatsian index” (YYI) against year, as shown in Figure 50.5. A straight line of best fit (produced by the least squares method) is drawn on this plot, from which the date of unseen texts can be read off. The “youthful Yeatsian index” was defined as follows:

$$YYI = (YY - OY)/(YY + OY)$$

To date an unseen text, one should count the number of substrings distinctive of either the younger or older Yates (these are YY and OY respectively), then calculate YYI. Find this value of YYI on the vertical axis, then read across to the line of best fit, then read down to find the estimated date of authorship on the horizontal axis.

5. Conclusion

In this review of computational stylometry we have looked mainly at studies of disputed authorship, but also considered that very similar techniques have been used for studies of genre, gender, and variation within a single author's writings which occur with time or between characters. The basic technique is to find samples of text known to have been written by each author considered in a particular study. In these texts we must find, whether manually or using machine-learning techniques, features such as whole words or character sequences which act as discriminators, since they occur more frequently in texts of one category than another. The frequencies of these discriminators are then found for unseen texts. Statistical tests, or models based on biological analogy, are then used to convert the raw data of discriminator counts in each of the texts into a definitive answer as to who was the likeliest author of each unseen text.

In the introduction, a caveat was given that differences in the styles of individual authors can be swamped by differences in genre and time. In fact, many of the techniques described in this article are also used in studies designed specifically to look at genre and diachronic studies. A technique very similar to that of Forsyth (1999) was used by Milton (1998) to compare the English used by native speaker students and learners of English in Hong Kong. Instead of looking for sequences of up to eight characters to identify either the younger or older Yeats, he identified sequences of four words which were typically overused or underused by the learners of English compared with the native speakers. Recently, Baroni/Bernardini (2006) used text categorisation techniques to compare original Italian texts and texts translated into Italian ("translationese") from other languages. The techniques studied in this article are largely language independent, apart from the need for separate lists of function words for each language. Hu/Williamson/McLaughlin (2005) are building a corpus for a diachronic study of Chinese, where the characteristics of Chinese written at widely different times and in different genres will be sought.

6. Acknowledgements

Permission to reproduce Figure 50.1 in this article, which originally appears in the *Journal of the Royal Statistical Society*, was granted by Blackwell Publishing. This figure was Figure 1 of Holmes (1992). Permission to reproduce Figures 50.2 to 50.5 in this article, which originally appeared in the *Journal of Literary and Linguistic Computing*, was kindly granted by the Oxford University Press. These figures were Figure 5 of Hilton and Holmes (1993), Figure 11 of Holmes/Gordon/Wilson (2001), Figure 1 of Merriam/Matthews (1993), and Figure 1 of Forsyth (1999). All figures appear by kind permission of the original authors.

7. Literature

Antosch, F. (1969), The Diagnosis of Literary Style with the Verb-Adjective Ratio. In: Dolezel, L./Bailey, R. W. (eds.), *Statistics and Style*. New York: American Elsevier.

Baayen, H. H./van Halteren, H./Tweedie, F. (1996), Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution, In: *Literary and Linguistic Computing* 11, 121−132.

Baroni, M./Bernardini, S. (2006), A New Approach to the Study of Translationese: Machine Learning the Difference between Original and Translated Text. In: *Literary and Linguistic Computing* 21, 259−274.

Barr, G. K. (1997), The Use of Cumulative Sum Graphs in Literary Scaleometry. In: *Literary and Linguistic Computing* 12(2), 105−111.

Biber, D. (1995), *Dimensions of Register Variation*. Cambridge: Cambridge University Press.

Binongo, J. N. G./Smith, M. W. A. (1999), The Application of Principal Component Analysis to Stylometry. In: *Literary and Linguistic Computing* 14(4), 445−465.

Brunet, E. (1978), *Vocabulaire de Jean Girandoux: Structure et évolution*. Paris: Slatkine.

Burrows, J. (2002), 'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship. In: *Literary and Linguistic Computing* 17(3), 267−287.

Campbell, D. (1992), Writing's on the Wall. In: *The Guardian*, October 8, 1992.

Campbell, D. (1997), Body of Evidence. In: *The Guardian*, August 7, 1997.

Canter, D. (1992), An Evaluation of the "Cusum" Stylistic Analysis of Confessions. In: *Expert Evidence* 1(3), 93−99.

Collins, J./Kaufer, D./Vlachos, P./Butler, B./Ishizaki, S. (2004). Detecting Collaborations in Text. In: *Computers and the Humanities* 38, 15−36.

DeForest, M./Johnson, E. (2001), The Density of Latinate Words in the Speeches of Jane Austen's Characters. In: *Literary and Linguistic Computing* 16(4), 389−401.

Dixon, P./Mannion, D. (1993), Goldsmith's Periodical Essays − a Statistical Analysis of Eleven Doubtful Cases. In: *Literary and Linguistic Computing* 8(1), 1−19.

Forsyth, R. S. (1999), Stylochronometry with Substrings, or: A Poet Young and Old. In: *Literary and Linguistic Computing* 14(4), 467−477.

Francis, I. (1966), An Exposition of a Statistical Approach to the Federalist Dispute. In: Leed, J. (ed.), *The Computer and Literary Style*. Kent OH: Kent State University Press.

Hilton, M. L./Holmes, D. I. (1993), An Assessment of Cumulative Sum Charts for Authorship Attribution. In: *Literary and Linguistic Computing* 8, 73−80.

Holmes, D. I. (1991), Vocabulary Richness and the Prophetic Voice. In: *Literary and Linguistic Computing* 6, 259−268.

Holmes, D. I. (1992), A Stylometric Analysis of Mormon Scripture and Related Texts. In: *Journal of the Royal Statistical Society Series A* 155, 91−120.

Holmes, D. I. (1997), Stylometry, its Origins, Development and Aspirations. In: Rudman, J./Holmes, D. I./Tweedie, F. J./Baayen, R. H. (chairs), session entitled The State of Authorship Attribution Studies. In: *ACH-ALLC '97 Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computers*. Kingston, Ontario, Canada, June 3−7, 1997. http://www.cs.queensu.ca/achallc97/papers/s004.html.

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

Holmes, D. I./Gordon, L. J./Wilson, C. (2001), A Widow and her Soldier: Stylometry and the American Civil War. In: *Literary and Linguistic Computing* 16(4), 403−420.

Holmes, D. I./Tweedie, F. (1995), Forensic Stylometry: A Review of the Cusum Controversy. In: *Revue Informatique et Statistique dans les Sciences Humaines* 31, 19−47.

Honoré, A. (1979), Some Simple Measures of Richness of Vocabulary. In: *Association for Literary and Linguistic Computing Bulletin* 7(2), 172−177.

Hoorn, J. F./Frank, S. L./Kowalczyk, W./van der Ham, F. (1999), Neural Network Identification of Poets Using Letter Sequences. In: *Literary and Linguistic Computing* 14(3), 311−338.

Hoover, D. I. (2002), Frequent Word Sequences and Statistical Stylistics. In: *Literary and Linguistic Computing* 17(2), 157−180.

Hoover, D. I. (2003), Frequent Collocations and Authorial Style. In: *Literary and Linguistic Computing* 18(3), 261−286.

Hoover, D. I. (2004), Another Perspective on Vocabulary Richness. In: *Computers and the Humanities* 37(2), 151−178.

Hu, X./Williamson, N./McLaughlin, J. (2005). Sheffield Corpus of Chinese for Diachronic Linguistic Study. In: *Literary and Linguistic Computing* 20, 281−293.

Jackson, MacD. P. (2002), Pause Patterns in Shakespeare's Verse: Canon and Chronology. In: *Literary and Linguistic Computing* 17(1), 37−46.

Kenny, A. J. P. (1982), *The Computation of Style*. Oxford: Pergamon Press.

Kjell, B. (1994), Authorship Determination Using Letter Pair Frequency Features with Neural Network Classifiers. In: *Literary and Linguistic Computing* 9, 119−124.

Koppel M./Argamon, S./Shimoni, A. R. (2002), Automatically Categorizing Written Texts by Author Gender. In: *Literary and Linguistic Computing* 17(4), 401−412.

Laan, N. (1995), Stylometry and Method. The Case of Euripides. In: *Literary and Linguistic Computing* 10(4), 271−278.

McEnery, A. M./Oakes, M. P. (2000), Authorship Identification and Stylometry. In: Dale, R./Moisl, H./Somers, H. (eds.), *Handbook of Natural Language Processing*. New York: Marcel Dekker, 545−562.

Mealand, D. L. (1995), Correspondence Analysis of Luke. In: *Literary and Linguistic Computing* 10, 85−98.

Merriam, T. V. N./Matthews, R. A. J. (1993), Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Marlowe. In: *Literary and Linguistic Computing* 9(1), 1−6.

Michaelson, S./Morton, A. Q. (1976), Things ain't What they Used to be. In: Jones, A./Churchhouse, R. F. (eds.), *The Computer in Literary and Linguistic Studies*. Cardiff: The University of Wales Press, 79−84.

Milić, L. T. (1966), Unconscious Ordering in the Prose of Swift. In: J. Leed (ed.), *The Computer and Literary Style.* Kent OH: Kent State University Press, 79−106.

Milton, J. (1998), Exploiting L1 and Interlanguage Corpora in the Design of an Electronic Language Learning and Production Environment. In: Granger, S. (ed.), *Learner English on Computer*. Harlow: Longman, 186−198.

Morton, A. Q. (1978), *Literary Detection*. East Grinstead: Bowker Publishing.

Morton, A. Q./Michaelson, S. (1990), *The Qsum Plot*. Technical Report CSR-3-90, University of Edinburgh.

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

Pollatschek, M./Radday, Y. T. (1985), Vocabulary Richness and Concentration. In: Radday, Y. T./Shore, H. (eds.), *Genesis − An Authorship Study*. Rome: Biblical Institute, 191−214.

Rayson, P./Leech, G./Hodges, M. (1997), Social Differentiation in the Use of English Vocabulary: Some Analyses of the Conversational Component of the British National Corpus. In: *International Journal of Corpus Linguistics* 2(1), 133−152.

Santini, M. (2004). A Shallow Approach to Syntactic Feature Extraction for Genre Classification. In: *7th Annual Research Colloquium for the UK Special Interest Group for Computational Linguistics (CLUK)*. University of Birmingham, 207−214.

Sichel, H. S. (1975). On a Distribution Law for Word Frequencies. In: *Journal of the American Statistical Association* 70, 542−547.

Tweedie, F. J./Holmes, D. I./Corns, T.N (1998), The Provenance of De Doctrina Christiana, Attributed to John Milton: A Statistical Investigation. In: *Literary and Linguistic Computing* 13(2), 77−87.

Waugh, S./Adams, A./Tweedie, F. (2000), Computational Stylistics Using Artificial Neural Networks. In: *Literary and Linguistic Computing* 15(2), 187–197.

Woods, A./Fletcher, P./Hughes, A. (1986), *Statistics in Language Studies.* Cambridge: Cambridge University Press.

Yang, Y./Pedersen, J. (1997), A Comparative Study on Feature Selection in Text Categorization. In: *International Conference on Machine Learning (ICML-97)*. Nashville, TN, 412–420.

Yardi, M. R. (1946), A Statistical Approach to the Problem of Chronology in Shakespeare's Plays. In: *Sankhya (Indian Journal of Statistics)* 7(3), 263–268.

Yule, G. U. (1944), *The Statistical Study of Literary Vocabulary*. Cambridge: Cambridge University Press.

Michael P. Oakes, Sunderland (UK)