Authorship Attribution

David I. Holmes

Department of Mathematical Sciences, University of the West of England, Coldharbour Lane, Bristol BS16 1QY, UK
e-mail: di_holmes@csd.uwe.ac.uk

Abstract: This paper considers the problem of quantifying literary style and looks at several variables which may be used as stylistic "fingerprints" of a writer. A review of work done on the statistical analysis of "change over time" in literary style is then presented, followed by a look at a specific application area, the authorship of Biblical texts.

Key Words: stylometry, authorship, vocabulary, model, multivariate

1. Introduction

The statistical analysis of a literary text can be justified by the need to apply an objective methodology to works which for a long time may have received only impressionistic and subjective treatment. Hesitation by literary scholars and mistrust of such a blatantly quantitative approach may be alleviated by choosing the least contestable mode of analysis, namely that of counting. The stylometrist therefore looks for a unit of counting which translates accurately the "style" of the text, where we may define "style" as a set of measurable patterns which may be unique to an author. The advent of the computer has meant that data for this purpose is now readily available in the form of a concordance or word-index to a literary work. The choice of the number of different words (types) in a text as a counting unit allows the stylometric analyst the freedom of working on the raw data and of operating a lemmatization according to norms which he can define himself, for example, the subsuming of singular and plural forms. This choice may run the risk of treating the individual written or printed word as unduly sacrosanct, yet, to date, no stylometrist has managed to establish a methodology which is better able to capture the style of a text than that based on lexical items.

As soon as the researcher starts counting words, especially with the help of a computer to index them, questions are inevitably asked about the relationship of the number of different words in a text (V) to the text length (N). Does the relationship V/N vary according to the length of text analysed, for example? A sense of the variety and richness in a text can be obtained by establishing the presence of rare or technical words, that is by studying peripheral or marginal vocabulary. Just as important, however, and what may readily go unnoticed, is the usage made of everyday, non-contextual, function or "filler" words drawn from mainstream vocabulary.

Computers have been widely used in attempts to establish the authorship of anonymous or doubtful texts. Writing in a forensic context Bailey (1979) proposed three rules to define the circumstances necessary for authorship attribution:

(i) the number of putative authors should constitute a well-defined set;

David Holmes is a Principal Lecturer in Statistics at the University of the West of England, Bristol with specific responsibility for co-ordinating the research programmes in the Department of Mathematical Sciences. He has taught literary style analysis to humanities students since 1983 and has published articles on the statistical analysis of literary style in the Journal of the Royal Statistical Society, History and Computing, *and* Literary and Linguistic Computing. *He presented papers at the ACH/ALLC conferences in 1991 and 1993.*

Computers and the Humanities **28**: 87–106, 1994.

(ii) the lengths of the writings should be sufficient to reflect the linguistic habits of the author of the disputed text and also those of each of the candidates;
(iii) the texts used for comparison should be commensurate with the disputed writing.

A statistical study of disputed authorship should involve comparisons of the disputed text with works by each of the possible candidate authors using appropriate statistical tests on quantifiable features of the texts, features which reflect the "style" of the writing as defined above. Bailey lists the general properties for such features:

> They should be salient, structural, frequent and easily quantifiable, and relatively immune from conscious control.

By measuring and counting such stylistic traits we hope to discover characteristics of a particular author and to distinguish genuine differences of style from chance variations in usage. This paper discusses criteria which may serve as a basis of measurement, followed by a review of work done on the authorship of biblical texts. Coverage is not claimed to be exhaustive, rather it is intended to impart a flavour of what the science of stylometry is about and to highlight its rapid progress during the past two decades.

2. Word-Length

Broadening an earlier suggestion advanced by Augustus de Morgan, Mendenhall (1887) proposed that word-length might be a distinguishing characteristic of writers. Mendenhall took several authors and constructed frequency distributions of word-length, showing that texts with the same average word-length might possess different spectra. He studied the Shakespeare/Bacon controversy and discovered that "in the characteristic curve of his plays, Christopher Marlowe agrees with Shakespeare about as well as Shakespeare agrees with himself."

Brinegar (1963) adopted the word-length approach to show that Mark Twain did not write *The Quintus Curtius Snodgrass Letters*, and Mosteller and Wallace (1964) tested Mendenhall's method in their study of the authorship of *The Federalist Papers*. Both these studies used χ^2 tests to compare distributions, a technique based on the unproven premise that an author's writings are random samples from his/her own fixed frequency distribution of word-lengths. The context-dependence of vocabulary means that this, at best, can be only an approximation.

Comparatively recent work by Smith (1983) indicates that when works which are of various literary genres are compared (or works written during different eras), the differences observed are likely to exceed greatly any distinguishing characteristics which may reliably identify authors. Furthermore, when works in the same literary form by different contemporaneous authors are compared, their word-length distributions may appear so similar that they seem to have been written by the same hand.

Smith concludes:

> Mendenhall's method now appears to be so unreliable that any serious student of authorship should discard it.

3. Syllables

W. Fucks (1952), in a pioneering work on both English and German authors, calculated the average number of syllables per word, the relative frequencies of the i-syllabled words, and their distribution in the text (i.e. the length of the gaps between i-syllabled words). Some of his conclusions point towards certain distinctive traits of an author, and various genres (poetic or prose) may also be identified quantitatively on the basis of such characteristics. In a later study, Fucks and Lauter (1965) discovered that frequency distributions of syllables per word discriminated different languages more than specific authors.

Bruno (1974) built upon the work of Fucks by using syllable averages from frequency distributions of i-syllabled words in a study examining heterogeneity within the medieval German epic *Nibelungenlied*, the reasoning being that row and column means of this two-dimensional frequency distribution would serve as a more sensitive stylistic discriminator.

A detailed study of the distribution of syllables per word in English texts has been undertaken by Brainerd (1974). He found that a model based on a translated negative binomial distribution was a better fit to such distributions than Fucks' translated Poisson distribution. Brainerd carefully considered whether or not the number of syllables in a pair of consecutive words may be viewed as being independent of each other since all his

samples are samples of consecutive words. He found that independence obtained to a large extent between consecutive words but suggested that there may be gross changes in the overall distribution of syllable counts as style changes from one kind of writing (e.g. narrative) to another (e.g. conversation). Brainerd concludes that some authors' styles are more homogeneous than others as regards syllable count and it would appear that the distribution of syllables per word in a corpus, being an easily accessible index of its style, is one area that may prove profitable in authorship studies.

4. Sentence-Length

Yule (1938) suggested sentence-length as a method for determining authorship and used this approach in a study of the disputed authorship of *The Imitation of Christ*. He concluded that sentence-length statistics are not a wholly reliable indicator in matters of this sort but he did raise some important questions concerning the definition of "sentence" for statistical work. Williams (1940) discovered that by taking a frequency distribution of the logarithm of the number of words per sentence, an approximation to a Normal distribution was found for each author. The lognormal model suggested by Williams and used by Wake (1957) must be rejected on several grounds, one of which is that most of the observed sentence-length distributions, after logarithmic transformation, are negatively skew. Sichel (1974) suggests a compound Poisson distribution for representing sentence-length distributions which appears to work well. Morton (1965) also used sentence-length for tests of authorship of Greek prose but little further experience has been reported in studies of this type although Kjetsaa (1979) did use sentence-length in his impressive study of authorship of *The Quiet Don*.

The disadvantages of using sentence-length as a stylistic variable are that it is under the conscious control of an author and, since division into sentences depends upon the punctuation, it is valid only to compare texts which preserve either the author's punctuation or are edited by one person. Smith (1983) compares sentence-length analysis with known facts of authorship in three separate studies and concludes:

> The statistics of sentences in each case confirmed the outcome derived by other methods. In each of the three tests, however, the information provided is not sufficiently strong to warrant the use of such measures as a stand-alone technique to discriminate between authors.

Both Kjetsaa and Tallentire (1972) agree that summary measures such as average sentence-lengths are of little use in authorship studies but distributions of sentence-lengths can be useful, even on their own.

5. Distribution of Parts of Speech

Significant stylistic traits may be found by noting the different percentages of nouns, verbs, adjectives, adverbs and other parts of speech in a text, if they can be defined accurately. Somers (1966) suggests:

> A more cultivated intellectual habit of thinking can increase the number of substantives used, while a more dynamic empathy and active attitude can be habitually expressed by means of an increased number of verbs. It is also possible to detect a number of idiosyncrasies in the use of prepositions, subordinations, conjunctions and articles.

One might also want to study the proportion of certain classes to others. Antosch (1969) studies the behaviour of the verb-adjective ratio for various literary genres and shows that this ratio is dependent on the theme of the work, e.g. folk tales have high values, scientific works have low values.

Brainerd (1974) produces a meticulous and intensive study into whether or not article frequency could be used as a style indicator. Carefully checking whether a sample of consecutive words may be considered random in regard to the occurrence of articles, Brainerd finds evidence of an overall Poisson character in the works sampled. He concludes that article counts in blocks of text appear not to be author-specific but they do appear to be sensitive to differences in kinds of writing. Newspaper writing, narrative and formal expository writing tend to employ more articles in general than novels, correspondence and autobiography. He also finds evidence of a connection between the number of articles and the number of pronouns in a text.

In a separate article, Brainerd (1973) uses article and pronoun analysis in an attempt to distinguish between a novel and a romance, such

parts of speech being particularly sensitive to variation in the degree of formality of writing. He found that novels tend to give higher pronoun counts and smaller article counts than romances, but also found a definite lack of homogeneity among romance authors. This latter discovery is hardly surprising since heterogeneity is in the nature of the romance genre.

It would appear from Brainerd's work that article and pronoun analysis would be interesting areas to examine in stylometric studies.

6. Function Words

Word-usage offers a great many opportunities for discrimination. Some words vary considerably in their rate of use from one work to another by the same author, others show remarkable stability within an author. For discrimination purposes we need context-free or "function" words to be able to conduct reliable comparisons between literary works.

Ellegard (1962) used function word frequencies in his study of the authorship of the *Junius Letters*, a series of political pamphlets written in 1769–1772 consisting of some 150,000 words of text. Ellegard obtained an ordered list of words which were positively or negatively characteristic of Junius in the sense that the writer used them more frequently or less frequently than his contemporaries. He calculated a "distinctiveness ratio" for each of these words, this ratio being given by

$$\text{distinctiveness} = \frac{\text{relative frequency of the word in Junius}}{\text{relative frequency of the word in a million-word sample of non-Junian writings.}}$$

Ellegard also examined the choices made by Junius from pairs of synonyms and it may be a characteristic feature of an author's style that whenever he/she has occasion to use a word with a particular meaning, one rather than the other of two synonymous ways of expressing the meaning is chosen.

Mosteller and Wallace (1964) used synonym preference in their impressive study of the authorship of *The Federalist Papers*. These were published anonymously in 1787–88 by Alexander Hamilton, John Jay and James Madison to persuade the citizens of New York to ratify the Constitution. Synonym-pairs were few in number, however, so they looked for words used by Hamilton and Madison with comparable frequency but at different rates. Function words such as prepositions, conjunctions and articles provided suitable instances and Mosteller and Wallace, using such discriminators, employed numerical probabilities to express degrees of belief about propositions such as "Hamilton wrote paper No.52" and then used Bayes' theorem to adjust these probabilities for the evidence in hand. The differences between the Ellegard and Mosteller-Wallace methods lie in the use of more powerful statistical methods by the latter, who saw their work as one application of statistical discrimination methods rather than as a means of settling the "Federalist" dispute.

Morton (1978) has developed techniques of studying the position and immediate context of word-occurrences or other words with which the writer tends to collocate it. His method consists essentially of applying a number of tests of words in prescribed positions in sentences, in collocations and as proportional pairs, to samples of the text whose authorship is in doubt and also to control texts, each test giving rise to a 2×2 contingency table. Smith (1983) suggests alternative approaches to handling this type of data using contingency tables of dimensions larger than 2×2.

Since writing this paper, however, Smith has become disillusioned with Morton's method and in two 1985 papers he demonstrates that it cannot distinguish reliably between the work of Elizabethan and Jacobean playwrights. The main weakness in the development of Morton's method is, according to Smith, a lack of technique in the acquisition of data, Morton's selection of features suitable for testing being based purely on a personal judgment of which features were likely to depend least upon context. Smith expresses doubt concerning the rigour of some of Morton's tests due to his failure to present comprehensively the data upon which his conclusions are based and on the small size of his samples.

In a 1987 study of the authorship of *Pericles*, Smith presents a method based on frequently occurring words. He restricts himself to first words of speeches and then conducts a complementary investigation using frequently occurring words

which do not occur as first words of speeches. A set of words to discriminate between candidate authors is initially derived and then subsets are chosen to distinguish between these dramatists taken in pairs. Proper names, verbs and pronouns are excluded. The criterion for selection of a word was that its rate of occurrence must be at least three per thousand in any of the plays under investigation. The plays were compared with each other by forming size 2×47 contingency tables, and then if a particular word contributed a value of 3.84 or more to the chi-square value it was chosen for the subset. The method revealed that Wilkins is more likely than Shakespeare to be the main author of Acts I and II of *Pericles*. Further work of this nature has been conducted by Smith in his 1991 study of the authorship of *The Raigne of King Edward the Third*.

Burrows (1987) presents a detailed and exhaustive study of Jane Austen's six published novels using the thirty most common words and making no distinction between function-words and content-words. These thirty words account for some 40% of the total text size in every case. Burrows shows that within each of three formal divisions-pure narrative, character narrative and dialogue-the six novels show roughly similar frequency patterns. This pattern is disrupted, however, when comparisons are made between these different categories. Burrows then breaks each pure narrative into a number of overlapping successive segments and compares them with each other and with the corresponding segments of another narrative. Using a correlation matrix approach he is able to compare and contrast Jane Austen with other authors, indicating a genuine authorial difference in narrative style and even between the narrative styles of Jane Austen at different stages of her career. This "change over time" aspect of literary style will be considered in a later section of this paper.

The conjecture that the distribution of function words may be a general fingerprint for authorship has been questioned by Damerau (1975), who used a computer to look for words occurring at least five times per 10,000 words in large samples of *Vanity Fair* and three American novels. Damerau's tests were based on the assumption that a word will be independent of context (and therefore a function word) if its occurrences follow a Poisson distribution. He found considerable diversity among his authors, words having a Poisson distribution for one author are widely divergent from such a distribution in another. For some authors many words are distributed according to a Poisson distribution, while for other authors only a few words are so distributed.

Oakman (1980) comments:

> The lesson seems clear not only for function words but for authorship word studies in general: particular words may work for specific cases such as *The Federalist Papers* but cannot be counted on for other analyses.

This seems unwarranted since Damerau's limited samples (only *Vanity Fair* had more than one sample taken from it) fail to show whether or not the phenomenon is characteristic of an author rather than of the work sampled. More studies are needed here to investigate stability within an author's work and between works by the same author.

7. The Type-Token Ratio

Tallentire (1973) asserts that:

> No potential parameter of style below or above that of the word is equally effective in establishing objective comparison between authors and their common linguistic heritage.

He points out that the lexical level is the obvious place to initiate stylistic investigations, since questions about style are essentially comparative and more data exist at the lexical level than at any other in the form of computed concordances. As Gregory (1964) suggests, analysis at syntactic and higher levels usually resorts to lexical analysis when semantic problems are encountered.

One of the fundamental notions in stylometry is the measurement of what is termed the "richness" or "diversity" of an author's vocabulary. The basic assumption is that the writer has available a certain stock of words, some of which he/she may favour more than others. If we sample a text produced by the writer, we might expect the extent of his/her vocabulary to be reflected in the sample frequency profile. If, furthermore, we can find a single measure which is a function of all the vocabulary frequencies and which adequately characterizes the sample frequency distribution we may then use that measure for comparative purposes. A simple single measure,

which is also computed under the "statistics" option in the Oxford Concordance Program, is the type-token ratio

If N = the number of units (word occurrences) which form the sample text (tokens), and
V = the number of lexical units which form the vocabulary in the sample (types),
then the type-token ratio is defined by R = V/N.

As has frequently been pointed out in the literature, applications of the index R are severely limited by its lack of stability with respect to variation in sample size. While N may increase without bound, the total vocabulary will in practice be likely to be finite. Kjetsaa (1979) restricted himself to studying the behaviour of R with N fixed at 500 in his investigation of *The Quiet Don* and found an approximate Normal distribution of types per 500 tokens in all texts analysed. Certainly it would seem that the type-token ratio would only be useful in comparative investigations where the value of N is fixed.

Baker (1988) defines the inverse of the type-token ratio as "pace," i.e. the rate at which new words are generated by an author. In a study of the works of Marlowe and Shakespeare, Baker claims that the pace of a text is "extremely characteristic of an author's style" and is independent of both text length and genre. He also contends that pace is related to the maturity and development of a writer. His work, however, has no statistical theory to back up this contention and is based on simple word counts of these two Elizabethan authors, most of whose works are in any case similar in length. His most extreme result happens to be the work of smallest text size.

8. Simpson's Index (D)

Simpson (1949) suggested a different approach to the measurement of diversity based on the chance (D) that the two members of an arbitrarily chosen pair of word tokens will belong to the same type. To calculate D, we simply divide the total number of identical pairs in the sample by the number of all possible pairs, i.e.

$$D = \sum_i r(r-1)V_r/\{N(N-1)\}, \quad (r = 1, 2 \ldots),$$

where V_r = number of types which occur just r times in a sample of text. D is in practice much more tractable than r since it does not depend on V_0, the number of unobserved words.

Johnson (1979) shows that

$$E(D) = \sum_i p_i^2 \quad (i = 1, 2, \ldots)$$

that is, the probability that any two items chosen at random from the population of vocabulary words will belong to the same type. Hence D is not only an indicator of the diversity or richness of the sample vocabulary, but also it is an unbiased estimator of the corresponding population value, irrespective of the sample size N.

Recent work by Thoiron (1986) has, however, cast doubt on the suitability of this index as a measure of vocabulary richness. Experiments in which he increased the repetitivity of a text sometimes lead to higher, sometimes to lower values of D. He found that D was more sensitive to variations on the higher frequency words, which are nearly always function words, and that these carry excessive weight in the determination of D. The almost insignificant part attributed to the lower frequency words is unfortunate, since these make a strong contribution to the concept of vocabulary richness.

9. Yule's Characteristic (K)

Yule (1944) devised a well-known "Characteristic K," a measure of vocabulary richness based on the assumption that the occurrence of a given word is based on chance and can be regarded as a Poisson distribution. An unbiased estimator of $\sum p_i^2$ under the Poisson assumption is:

$$\sum_r r(r-1)V_r/N^2$$

from which

$$K = 10^4(\sum_r r^2V_r - N)/N^2$$

(NB 10^{-4} K = D(1 − 1/N))

The underlying Poisson assumption used by Yule in his formulation of K has come under attack and led to a reformulation of K by Herdan (1955) defined as a "coefficient of variation of a sample mean," Johnson considers that this does not offer any useful advantages over either K or D.

Yule's K has been surprisingly neglected in recent years although Bennett (1969) uses it on common nouns only in an investigation into vocabulary richness in *Julius Caesar* and *As You Like It*. Bennett found differences between Acts, reflecting the subject matter, but the plays themselves showed homogeneity.

Tallentire (1972) found that a wide range of K values is usually obtained when works are sampled and suggests that K is not well suited to attribution problems. He recommends that all words should be used when calculating K, not just common nouns, and that K values should not be used in isolation in authorship studies since the degree of repetition revealed in vocabulary is probably not an unconscious aspect of style. Sichel (1986) proves that under the Poisson assumption that the occurrence of a given word is based on chance, K is constant with respect to N, and Weitzman (1988) also comments on this favourable property of K.

10. Entropy

The expression for the entropy of a system (vocabulary) is

$$H = -\sum_i p_i \log p_i,$$

where p_i is the probability of appearance of the ith word type (found by dividing the number of occurrences of that word by the total number of words in the text). This variable is based on a thermodynamic concept of a literary text, namely, that with an increase in internal structure entropy decreases and with an increase in disorder or randomness the measure of entropy increases. Since the value will change according to how much text is analysed, the formula may be refined in order that works of different length may be compared.

Using

$$H = -100\sum_i p_i \log p_i / \log N$$

absolute diversity for any length text is measured as 100 while absolute uniformity remains zero.

Bruno (1974) comments:

> Only with the establishment of a sound methodology does it seem appropriate and promising to utilize statistical entropy as a possible discriminator of style.

This point is taken up by Johnson (1979) who stresses the problems of interpretation of H when samples of text are used, while the total number of word types is unknown. Johnson points out the danger of using such a measure for other than very informal inferences about the underlying population and shows that its use in comparative studies involving two or more texts has little theoretical justification.

Johnson concludes:

> There seems to me no doubt that of the measures which have been seriously considered in the literature the most satisfactory indexes of diversity for vocabulary studies are those based on estimates of the repeat rate (i.e. D and K). The claim has already been made in their favour that, on empirical evidence, they are extremely robust with respect to variation in the sample size. The knowledge that they are also unbiased estimates of an easily interpreted population value, together with the added bonus of some associated sampling theory, in my opinion can only enhance their usefulness to those scholars who are concerned with the measurement of style and vocabulary.

11. Vocabulary Distributions: Models for Vocabulary Curves

Another way of measuring vocabulary is to count how many words are used once in a text, how many are used twice, and so forth. The advantage of this method is that the measurement is independent of context and thus can be applied to works on completely different subjects, the information revealing something about the attitude of an author towards rare words and a diversified vocabulary. When one considers the list of lexical units that appear in a text of a given length, it is seen that the number of words that are used once (hapax legomena) is greater than that of the words used twice (dislegomena), that the latter are more numerous than those appearing three times, and so on. The problem is to determine if this decrease obeys a law and this section reviews attempts to deal with this.

Mathematical models for the frequency distribution of V_i (the number of vocabulary items appearing exactly i times) have aroused the interest of statisticians for many years. Zipf (1932) was the first to reveal that a relationship exists between the number of occurrences (i) and their frequencies V_i. When the logarithms of occurrences are plotted against the logarithms of frequencies an approximate straight line obtains.

This "law" has been severely criticized, most notably by Herdan (1966), because the deviation from the straight line cannot be easily dismissed; too many occurrences, i.e. the tail of the distribution, are associated with one word only.

Yule (1944) conjectured that the correct distribution for word frequencies would be a compound Poisson distribution but his efforts to construct a model to account for the facts ended in failure. A superior fit can be obtained using what is known in statistical literature as the Waring distribution. In recognition of the pioneering work of Herdan (1964) in applying this distribution to the frequency of vocabulary items, the distribution has come to be known in linguistic literature as the Waring-Herdan model.

The Waring law predicts the development of an expression $(x - a)^{-1}$ in which $o < a < x$. By multiplying by $(x - a)$ the expression itself on the one hand, and the series which is obtained by its expansion, on the other, we obtain unity for the former and a decreasing series for the latter:

$$\frac{x-a}{a} + \frac{(x-a)a}{x(x+1)} + \frac{(x-a)a(a+1)}{x(x+1)(x+2)} + \ldots + \frac{(x-a)a(a+1)\ldots(a+n-2)}{x(x+1)(x+2)\ldots(x+n-1)}$$

of which the sum is one.

The hypothesis is that the term of rank *n* represents the probability of a lexical item appearing *n* times in the text. This term is then defined by the equation:

$$P_n = \frac{(x-a)a(a+1)\ldots(a+n-2)}{x(x+1)(x+2)\ldots(x+n-1)}$$

The x and a parameters must then characterize the length of the text, the extent of its vocabulary and the dispersion or richness of its vocabulary. The solution proposed by Herdan is:

$$a = \left[\frac{V}{V - V_1} - \frac{V}{N} - 1\right]^{-1}$$

and

$$x = \frac{aV}{V - V_1}$$

where V_1 = the number of lexical units which appear only once. From this, it follows that the expected number of words employed exactly n times in the text is VP_n.

Muller (1969) has shown that this model is reasonably successful for text lengths $100 < N < 100{,}000$, yet Herdan made a mistake in his calculations which was demonstrated by Dolphin (see Muller, 1975) who produced a corrected moment estimator

$$a = \left[1 - \frac{V_1}{V}\right]\left[\frac{N}{V} - 1\right]\left[\frac{NV_1}{V^2} - 1\right]^{-1}$$

Herdan's mistake gave Dolphin and Muller the ingenious idea of evaluating the total range of a writer's lexicon by calculating the number V_o of word-types with o token. They go on to obtain a probability-function on this lexicon and then obtain a model equivalent to the Waring-Herdan model using the revised estimates of x and a. The Waring-Herdan and Dolphin-Muller laws are generally considered the best existing models to fit vocabulary curves.

Yule's conviction of a compound Poisson law was taken up by Sichel (1975) who proposed a new family of compound Poisson distributions as a model for word frequency counts. Sichel obtained an excellent fit of his distribution to numerous word occurrence data from different texts but did not interpret his parameters. Pollatschek and Radday (1981) applied Sichel's distribution to some Hebrew Biblical texts, showing that the two parameters involved, α and θ, could be interpreted as measuring vocabulary richness and vocabulary concentration respectively.

They plotted various theoretical distributions of the Sichel model for different sets of values of α and θ, and noted that the slope of the tail of the distribution was defined by θ alone whilst that of the head by α alone. Pollatschek and Radday conclude that α and θ together describe the whole vocabulary distribution in full, and the Sichel distribution hence becomes a powerful tool in measuring the richness of a writer's vocabulary. Parameters α and θ are not negatively correlated, in fact any of the four (high, low) possibilities may characterize a text.

Sichel dealt with the problem, first noted by

Yule, of whether or not words of one particular kind only, e.g. nouns, should be considered in the vocabulary distribution. His examples mainly concerned nouns and he concluded:

> where word counts are performed on all types, including nouns, verbs, adjectives, adverbs, pronouns, prepositions and conjunctions, one should not be surprised to encounter anomalies which are inherent in the superimposition of entirely different statistical populations.

Pollatschek and Radday included all words, which they felt seemed only proper for Hebrew, but warned that, for English texts, some words such as the definite and indefinite articles are so predominant that their inclusion might cause an extremely long tailpiece bound to distort the estimation. They concluded:

> The Sichel distribution . . . appears to be a promising tool in studies of verbal behaviour in general and in enquiries into homogeneity and disputed single authorship in particular.

An application of this distribution will be discussed in the section which reviews work done on the authorship of biblical texts, later on in this paper.

Finally, Delcourt (1981) proposed both 2-parameter and 3-parameter log-linear models and obtained excellent fits for French texts in which all grammatical categories were taken into account. His models have yet to be applied to other languages and clearly much work remains to be done in the field of modelling vocabulary distributions.

12. Comparing Vocabulary in Texts of Different Lengths

The assessment of relative vocabulary richness when comparing the works of different authors is complicated by the effect of different text length. As N increases, V also increases but at a steadily diminishing rate. Muller (1964) proposed a means of scaling down a text to a smaller size without changing its structure, thereby eliminating the effect of text size.

The method is based upon the fact that if a text is reduced by a fraction u where $u = 1 - (N_f/N)$, the quantity N_f being the length of the text after reduction, then the probability of disappearance of a vocabulary item that had appeared only once is u, the probability of disappearance of a twice appearing word is u^2, and so forth. Thus, in the long run the number of vocabulary items that can be expected to disappear is $\sum V_i u^i$ $(1 < i < \infty)$, and the expected value of the final number of i vocabulary items that remain, V_f, is given by

$$V_f = \sum_{i=1}^{\infty} V_i(1 - u^i).$$

The application of this formula requires knowledge of the complete table of the frequency distribution of the vocabulary items appearing exactly i times, that is V_i, and it is possible to combine the text reduction procedure of Muller with the Waring-Herdan distribution of vocabulary item frequency. We therefore have the means of evaluating the expected number of different words in a text of any size whose words are drawn at random from a corpus of words having the frequency distribution V_i.

Combining Muller's procedure with the Waring-Herdan formula reveals that V_f/V is a function only of u, N/V and V_1/V. Ratkowsky and Hantrais (1975) have computed tables listing V_f/V for various values of these three ratios which show that either for large V_1/V or low u, the fraction V_f/V is virtually independent of N/V. The tables naturally depend on the Waring-Herdan distribution providing a good fit to V_i and, although it sometimes has a tendency to overestimate the lower frequencies and to underestimate the higher frequencies, the table should nevertheless make it possible to calculate which of two texts of different sizes has the richer vocabulary and contribute towards establishing an objective norm for vocabulary richness.

An alternative to using tables is to use Muller curves from which one can read the expected vocabulary size of a text from the text length. Ule (1982) uses Muller curves for Marlowe and Shakespeare in his work on Elizabethan English. For authorship attribution problems we may compare the *actual* number of different words in a text against what we would *expect* if the author was X, provided of course, we have the distribution V_i for author X. Muller's procedure does assume that an author composes his works by taking his words at random (without replacement) from all his published works (or some equivalent corpus).

An ingenious amendment to Muller's procedure has been proposed by Hubert and Labbe (1988). Their "partition model" is founded on the idea that, on the one hand, an author uses a general non-specialised vocabulary which contains, for example, articles, prepositions, pronouns and common verbs, whereas, on the other hand, the author draws upon several local or specialised vocabularies which encompass the terms used within a single excerpt only. The contributions made by the local vocabularies (V_s) and the general vocabulary (V_g) to the total vocabulary (V) is measured by a coefficient p:

$$p = \frac{V_s}{V}$$

and

$$q = 1 - p = \frac{V_g}{V} \quad (0 \leq p \leq 1)$$

The expected value of the final number of vocabulary items that remain after text reduction by a fraction u, may then be expressed as

$$V_f = puV + q[V - \sum_{i=1}^{\infty} V_i(1 - u^i)]$$

Hubert and Labbe label p as the "coefficient of vocabulary partition" and describe it as estimating an intrinsic character of a text, namely the division between the general vocabulary and the specialised vocabularies which have been drawn on to produce it. They proceed to derive a least squares estimator for p based on the differences between observed and expected values of V_f for various fractions of a text and then compare their results with Muller's original procedure applied to the works of Racine. Their partition model closely follows the observed values whereas Muller's formula consistently provides over-estimates of V_f.

Ule also proposes an ingenious measure known as Relative Vocabulary Overlap (RVO) to test whether 2 texts of different lengths are written by the same author. RVO depends on two quantities, namely Absolute Vocabulary Overlap (AVO) and Expected Vocabulary Overlap (EVO). To calculate the AVO we need alphabetical lists of word frequencies for each text, the AVO being the sum of the lower values (including zero) of the two frequencies with which each word (type) appears in the two frequency lists. The EVO is defined to be the number of words (tokens) that the two texts of length W_1 and W_2 words respectively, would have in common if the texts had been composed by drawing words at random (without replacement) from all the author's published work. The EVO is therefore a function only of W_1, W_2 and the vocabulary structure V_i of the author.

The RVO then equals AVO/EVO, which Ule, from his work on Marlowe and Shakespeare, finds to be a stable measure of vocabulary comparison.

13. Word Frequencies

Instead of using vocabulary distributions V_i where we count how many words occur i times, we could look at how many times individual words occur in the corpus under study. In his pioneering study, Zipf (1932) ranked the various words of a text according to decreasing frequency and plotted on log-log paper the ranks (r) against the corresponding number of times which the word of rank r occurred, obtaining a straight line configuration (Zipf's first law).

Tallentire (1972) discusses the difficulties of using word frequencies in authorship studies, pointing out that the bulk of any sample of written English is accounted for by the same few words recurring with the same relative frequency even in very different writings (10 per cent of the vocabulary of English providing 90 per cent of the text of all the volumes of literature in all our libraries). Tallentire illustrates the difficulty in detecting the person behind the language since not only do a relatively common set of highest-frequency words recur in much the same proportion in different authors, but common lowest-frequency words do also. He appeals for linguistic templates, i.e. norms for each literary period, genre and language so that we may isolate oddities peculiar to particular authors – an appeal that has largely remained ignored.

The distinction between ranked word frequencies and vocabulary distributions is analogous to the distinction between probability density and its cumulative distribution function. Mandelbrot (1959) derive one from the other.

14. Hapax Legomena

In vocabulary frequency distributions, the largest group is of words which occur once in the text, V_1, (hapax legomena) and it is natural that one measure at least of vocabulary richness should involve this particular count. Morton (1986) comments:

> The once-occurring words convey many of the elements thought to show excellence in writing, the range of a writer's interests, the precision of his observation, the imaginative power of his comparisons; they demonstrate his command of rhythm and of alternations. The potential of once-occurring words as an indicator of authorship seems obvious, as a group they display so much that appears characteristic of the individual and the choices which must be made in composition.

Morton defines a once-occurring word as a form not repeated in the sample and says that we ought not to concern ourselves with supplementing form by meaning (e.g. the word SAW could be either a noun or a verb), since the proportion of such occurrences is much less than 1% of the total according to detailed examinations of samples of both Greek and English. He does not, therefore, separate homographs.

Morton's theme in his 1986 paper is that it is the position of these once-occurring words in the sentence which enables one writer to be distinguished from another. Hapax legomena occur relatively infrequently as first words of sentences but at an enhanced rate as last words, and Morton goes on to study such patterns in the epistles of the Pauline corpus. The problem with this approach is that it involves punctuation in the defining of a sentence, a matter of lesser difficulty perhaps, in Greek prose, but somewhat dubious in cases like medieval texts or in Shakespeare where the punctuation often represents editorial intervention. Here Morton abandons punctuation and studies the positions of hapax legomena in relation to particularly frequent words such as A, AND, IN, etc. He claims that Shakespeare is entirely consistent in his habit of placing once-occurring words before and after such marker words.

Smith (1987a) conducts a detailed investigation into the validity of Morton's approach, examining both its statistical implications and the range and quality of his evidence. Smith concludes, correctly, that much of Morton's work is not sound from a statistical point of view, many calculations having been performed without regard for the existence of an underlying theory. There is no evidence, therefore, that once-occurring words in prescribed positions within sentences can discriminate between authors. The verdict was the same for Morton's collocations. This is not to say that hapax legomena are of little use in authorship attribution. Smith points out that Morton's tests depend only on numbers of once-occurring words in prescribed locations and not directly on their total in a text. The difficulty is simply that Morton has provided no real evidence to support his premise that authors are distinctive in how they place their rare words in sentences or in collocations.

The behaviour of hapax legomena has been investigated by Brainerd (1988). An hypothesis of a fixed maximum available vocabulary would imply that the expected number of hapax legomena would increase, assume a maximum and then decrease to 0 as $N \rightarrow \infty$. However, Brainerd shows that the relation between V_1 and V for a text of 200,000 words is very near a straight line with positive slope, and for the two-million-word Kierkegaard corpus $V_1/V = 0.4359$ whilst for a representative sample of 103 million words of American English $V_1/V = 0.4472$. Brainerd suggests that the vocabulary available to an author may not be fixed but may grow linearly with time. He also finds that, in the Kierkegaard corpus, for works with nearly equal N's the V's and V_1's can be quite different, indicating differing vocabulary use across the works.

A useful function involving hapax legomena has been suggested by Honoré (1979) and is defined by

$$R = \frac{100 \log N}{(1 - V_1/V)}$$

It directly tests the propensity of an author to choose between the alternatives of employing a word used previously or employing a new word. In the extreme case when each word-type in a text is used once only, $V_1 = V$, and R becomes infinite. When comparing texts therefore, the higher the R value the richer the vocabulary in the sense that a greater number of words appear infrequently. Honoré applied this formula to texts from 39 legal

authors, writing in Latin, contained in Justinian's *Digest* (A.D. 533) and found that it appeared stable above text sizes of N = 1,300 words. From an additional study of five works of Cicero, Honoré found that the first two works (dated to 70–66 B.C.) gave results in the 1,100s whilst the last three (dated to 45–43 B.C.) gave results in the 1,300s. If Cicero's vocabulary increased in richness over the intervening twenty year period, which seems plausible, then the R function may successfully measure change over time in vocabulary richness and be helpful when problems of dating are at issue.

15. Hapax Dislegomena

The behaviour of words used only twice in a given text (hapax dislegomena) has been investigated by Sichel (1986). Sichel found that the proportion of hapax dislegomena increased very rapidly with increasing N then stayed constant for a very long interval of token counts before dropping very gently towards zero as $N \rightarrow \infty$. His model showed stability of the proportion of hapax dislegomena between $1{,}000 < N < 400{,}000$ and tests on real data corroborated this theoretical result i.e. for a particular author, this proportion, as observed over a wide range of token counts, is virtually constant. As illustration, consider the following example taken from Sichel (1986):

	N	V_2/V
A sample of nouns	2000	0.1872
drawn from Macaulay's	4049	0.1767
Essay on Bacon	6004	0.1763
	8045	0.1792

Sichel postulates that whilst the proportion of hapax legomena is decreasing with increasing N, the near invariance of the proportion of hapax dislegomena may be due to an equilibrium being reached. With increasing N, as many words could be lost from this category to become thrice-occurring words as, proportionately, are received from the once-occurring word category.

This interesting discovery of the near constancy of the proportion of hapax dislegomena for a writer, whatever the number of tokens counted, suggests that the function V_2/V would be a very appropriate variable to use in studies of vocabulary richness.

16. Multivariate Studies

Multivariate statistical techniques such as factor analysis, discriminant analysis and cluster analysis have found increasing application in the analysis of literary style in recent years. An exhaustive review is not practicable here but some examples are given. Miles and Selvin (1966) applied factor analysis to the vocabulary of seventeenth century poetry and concluded that it was possible to determine trends of quantitative and qualitative influence of Petrarch, the Classics and the Bible on poets of that era. Somers (1966) effectively applied discriminant analysis to the works of Philo Alexandrus and to the collection of the Epistles of St Paul to test the assumption that Philo exerted some influence upon the *Letter to the Hebrews*. Bruno (1974) applied a stepwise discriminant analysis in order to derive a discriminating equation that would distinguish between nineteen high formulaic and twenty-two low-formulaic stanzas chosen from the *Nibelungenlied*, and examples of cluster analysis are to be found in articles by Bailey (1979) and Boreland and Galloway (1980).

Kjetsaa (1979) used a pool of stylistic variables, some of which were particularly appropriate for texts in Russian, in his meticulous study of *The Quiet Don*. His variables represented a spread of syntactic and vocabulary categories indicative of the broad basis of the study and in a later paper (1981), Kjetsaa used 15 parameters in order to detect Dostoyevsky's stylistic "fingerprints." A pool of 37 stylistic variables was also used by Ledger (1989) in his study of ancient Greek authors. These variables were the percentage occurrences of words containing specified letters, ending in specified letters or with the specified letter as the penultimate letter. Ledger's cluster analyses correctly differentiate these authors and he concludes:

> Ultimately, the justification of these methods is that it is clear that they work, and I would forecast that they will be extensively used in the future and will change considerably our approach to stylometry.

This rather begs the question as to why, in linguistic terms, an analysis using these variables should work at all – an inevitable question if only

because, in an inflected language like classical Greek, his "last letter" variables usually have direct grammatical significance.

Ledger uses discriminant analysis to study the status of disputed Platonic dialogues. Discriminant analysis is a suitable technique when there are a number of samples that can be initially divided into groups, e.g. by author. The method then classifies samples for which the true group is not known by determining which of the pre-defined groups most closely matches each unknown sample. The problem of Platonic authorship is not so simple and Ledger acknowledges that the way in which he uses discriminant analysis is unorthodox.

Principal components analysis is a statistical technique which has the advantage of requiring no underlying mathematical model. It aims to transform the observed variables to a new set of variables which are uncorrelated and arranged in decreasing order of importance. The principle aim, here, is to reduce the dimensionality of the problem and to find new variables which will help to make the data easier to understand. These new variables (or components) are linear combinations of the original variables and it is hoped that the first few components will account for most of the variation in the original data.

Holmes (1992) uses principal components analysis and cluster analysis in his investigation of the authorship of the *Book of Mormon*. Five variables are used, each measuring some aspect of vocabulary richness, and twenty-four textual samples drawn from the *Book of Mormon*, the *King James Bible*, *Doctrine* and *Covenants* and the personal writings of Joseph Smith are analysed. Holmes finds no evidence of multiple authorship within the *Book of Mormon*, the style of the Book being consistent with the style of the "prophetic voice" of Joseph Smith. Two variables which appear to play a major role in this analysis are Honoré's R statistic and the proportion of hapax dislegomena V_2/V, both of which are discussed earlier in this paper.

Holmes (1991) extends this work on vocabulary richness in an additional paper which incorporates the prophecies and personal writings of the English prophetess, Joanna Southcott. He shows that statistical procedures based on vocabulary richness prove sensitive enough to distinguish between samples from the same genre as well as being able to discriminate between samples from different genres.

Burrows and Hassall (1988) also use principal components analysis to distinguish between Henry and Sarah Fielding. For their variables, Burrows and Hassall compute the rate of occurrence of the fifty most frequent words in the texts, using disputed works and works for which there is no doubt as to whether Henry or Sarah was the author as their textual samples. Plotting the data in the space of the first two principal components appears clearly to assign the disputed texts to either Henry or his sister. In a later article Burrows (1992) gives more examples of this mode of analysis. He shows how *The Memoirs of a Lady of Quality* resembles Smollett in its word usage rather than fourteen of his contemporaries and, fascinatingly, how authors tend to group by era. Language, as measured by the ratio of occurrence of non-contextual function words, appears to have undergone a steady process of change, and gender analysis shows that clear differences between male and female authors during the eighteenth century become obliterated by the twentieth century.

Smith (1991b) uses the same technique in ascribing *The Revenger's Tragedy* to Middleton rather than Tourneur. It is clear that, with the power of the computer now readily available, to provide both word listings and statistical analyses of the nature of those discussed here, the selection of multiple criteria must now be favoured to tackle problems of authorship attribution.

17. Change Over Time in Literary Style

Closely related to issues of authorship attribution in the statistical analysis of literary style are the problems connected with specifying the sequence of composition of the works of a given author. Stylistic "fingerprints" do not always remain stable, perhaps a writer becomes less stylistically innovative after an early "peak of diversity" as certain words and patterns become increasingly preferred. To test this hypothesis Tallentire (1976) used what he called the "hapax/token ratio" which indicates what proportion of a story's vocabulary is used once only, on several stories by each of Virginia Woolf and Katherine Mansfield. This ratio is not independent of sample size and has therefore to be used with care, but for story groups

(having approximately equal token-totals) Tallentire found that hapax in later stories decrease suggesting lexical diversity decreases with age. The lexical data was readily available in concordances, and such word listings may be combined with the text reduction procedure of Muller to enable us to calculate hapax/token ratios which may be applied to texts of different sizes when studying the dependence of vocabulary richness on age. This is an exciting prospect and one which has not been taken up so far.

In cases in which the sequence of composition is unknown, analysts may look for trends that will enable them to place undated works of a given author. Cox and Brandwood (1959) attempted to place in order some of the works of Plato by studying the distribution of the last five syllables of each sentence, each syllable being classed as long or short. They found a marked difference between the distributions for the *Republic* and the *Laws* and the problem was to order the other works in decreasing order of affinity with the *Republic*, a work with a distribution similar to that of the *Republic* being deemed to have been written at about the same time.

The technique used was essentially discriminant analysis. A score

$$S_i = \log\left(\frac{\theta_{1i}}{\theta_{0i}}\right)$$

was assigned to each type of sentence ending where θ_{1i} and θ_{0i} represent the probabilities that a sentence-ending from the *Republic* and *Laws*, respectively, fall in the i'th category. In practice estimated scores were obtained by replacing such probabilities by sample percentage frequencies. If n_i is the percentage of sentence-endings in the i'th category in a Platonic work under study, then the mean score

$$m = \frac{1}{100}\sum_i n_i s_i$$

is used as a discriminator. Kemp (1976) points out that such discriminators may in general be used for purposes of chronological classification in addition to their use in problems of disputed authorship.

Cox and Brandwood use the mean scores to order the (undated) works of Plato yet this depends on their assumption that "Plato's change in literary style was monotone in time" (p.195), an assumption in support of which they offer no argument and which justifies scepticism of their conclusions.

During the latter half of the nineteenth century many students noted and commented on the slow but steady change in a particular aspect of Shakespeare's style, namely the number of lines in different classes such as broken lines, run-over lines, rhymed lines, etc.

Williams (1970), as a follow-up study, took ten samples of fifty consecutive lines of verse from each of the plays of Shakespeare and sorted the lines into three categories: those which ended with no pause, short pauses (as shown by commas or colons) and longer pauses. He found evidence of a slow increase in run-over lines as Shakespeare grew older, largely at the expense of lines with long pauses (full stops). Distinct differences were also found between the different types of plays commonly designated as Tragedies, Histories and Comedies but care needs to be taken here since most of the histories and light comedies are early works whilst tragedies are more concentrated in his later years.

Ule (1982) has attempted to ascertain the sequence of composition of seven plays by Christopher Marlowe. Ule states:

> It is typically the purely mechanical nonsemantic attributes of a text that appear to be immune to the influence of theme or genre and which tend to change slowly, perhaps monotonically rather than cyclically, during the lifetime of an author.

Ule uses nine such attributes, namely:

1) The percentage frequency of occurrence of i-letter words (i = 1 . . . 40).
2) The same for sentences up to 40 words in length.
3) The relative percentage frequency of the 40 most common words in Elizabethan plays.
4) The same for 40 prepositions.
5) The relative frequency of 40 function words, which every author must use but where he may show certain preferences.
6) The percentage frequency of occurrence of the 26 letters of the alphabet and of the 10 Arabic numerals.
7) The percentage frequency of words used once,

twice, etc., i.e. the vocabulary structure V_i ($i = 1 \ldots 40$).
8) The percentage frequency distribution of iV_i.
9) The relative vocabulary overlap, RVO.

He postulates that a pair of texts written one after another will have more closely matching frequency distribution curves or a higher RVO than two texts not written sequentially. Comparing his results with known sequences, Ule concludes that the textual parameters best able to detect the correct sequence are word-length distribution, connectives and RVO.

Bender and Briggum (1982) have studied the works of Joseph Conrad to try to detect a shift in style as the author developed. Conrad's collaborator, Ford Maddox Ford, claimed that Conrad's style changed as the two of them consciously developed a new movement, "Literary Impressionism", which Ford explains as being a mode of writing where the novelist tries to make the same impression on the reader that life makes on mankind – not a narration or report, but a series of disconnected impressions. The major alteration appears to be a shift from emphasis on action (physical verbs) to emphasis on the speaker's unique perception of that action (psychic and weak verbs). Bender and Briggum find such a lexical shift towards a higher ratio of psychic and weak verbs when comparing Conrad's middle style (*Lord Jim*) with his early style (*Almayer's Folly*) and they conclude with a justifiable appeal for concordances which are parsed. The traditional limitations of concordances are highlighted when studies require the isolation and classification of parts of speech.

18. The Authorship of Biblical Texts

The main contribution of statistics to religious studies has been in the area of the stylistic analysis of texts. It is important to stress, however, that the results from any statistical analysis should be combined with the findings of more conventional methods of biblical scholarship in order to provide the fullest possible explanation of the textual material under investigation.

Modern biblical scholarship is generally dated to the mid-nineteenth century when a thesis called the "documentary hypothesis" was developed. This held that several narrative strands could be discerned in the early Hebrew Bible, and that these strands could be attributed to different sources. The first suggestion that a question concerning disputed authorship of a biblical book might be settled statistically was made by Augustus de Morgan in 1851 (see Morton and McLeman (1980)), who proposed that the authorship of the *Letter to the Hebrews* could be investigated by an examination of sentence length. The formal use of statistical analysis to explore such hypotheses about the composition of biblical books did not, however, occur until the mid-twentieth century. The principal work in this area will now be reviewed, with an emphasis on the more recent studies.

The authorship of the New Testament *Pastoral Epistles* (Titus, I Tim., II Tim.) has long caused controversy and was investigated in 1959 by Grayston and Herdan, who looked at what they termed "vocabulary connectivity" between the Epistles. This measure involves the ratio

$$C = \frac{\text{(Words used once only in the given Epistle) + (Words common to all Epistles)}}{\text{Vocabulary of the given Epistle}}$$

as applied to all the Pauline Epistles. Grayston and Herdan found that most of the values of C were within the range 32–34%, but that two were outstanding: the value for the Pastorals (46%) and that for the Thessalonians (29%). The authors concluded that the linguistic evidence in terms of C was that the Pastorals showed less vocabulary connectivity with the total Pauline vocabulary than did the rest of the letters and that this was in full agreement with conclusions reached on scholarly grounds. These findings appeared to lend support to the hypothesis of non-Pauline authorship of the Pastorals. Grayston and Herdan then computed the bi-logarithmic type-token ratio (log V/log N) for all the Pauline Epistles and found that this value could be regarded as being "sensibly constant" except in the case of the three Pastorals. No significant tests were given, yet again the authors concluded that their findings confirmed in a formal way existing doubts about non-Pauline authorship for these Epistles. It may be noted, however, that no adjustments were made to allow for different text lengths.

Morton (1965) has also considered the authorship of the Pauline epistles, and his work has been reviewed above. As we have seen, it attracted considerable criticism and highlighted the problems of drawing valid inferences based on statistical characteristics of style even where such simple indicators as sentence length and word frequency are concerned. Morton found, using these criteria, that Romans, I and II Corinthians, and Galatians formed a relatively coherent group, but that other Epistles differed from these by significantly more than an examination of the variation within single Greek authors would lead us to expect. On this basis, Morton rejected the attribution of these other Epistles to Paul.

More recently Kenny (1986) has conducted a stylometric study of New Testament materials. In common with others, Kenny emphasises that any stylometric analysis must simply be regarded as new evidence which must be weighed in the balance along with the more traditional criteria applied by the biblical scholar. His analysis is based on 99 separate stylistic features each of which involved counts e.g. number of masculine nouns, number of neuter articles, total number of third-person pronouns. He concluded that John's Gospel was very unlike the Apocalypse, that Luke and Acts were close enough to be regarded as the work of the same author, and that the Pauline problem was not simply one of regarding only the first four Epistles as the work of a single hand. Kenny concedes that the most serious limitation of his study was the difficulty in working with short passages. This was why he had given statistics only for New Testament works as a whole, and not for samples within them.

Another influential school, that associated with Y T Radday, has applied statistical methods to authorship problems in the Hebrew Bible, in particular to books where biblical scholars have found evidence of possible compilation from several sources. This group of books comprises Isaiah (Radday, 1973), Zechariah (Radday and Wickmann, 1975), Judges (Radday, Wickman, Leb and Talman, 1977), Lamentations (Radday and Pollatschek, 1977) and Genesis (Radday, Shore *et al*, 1985). Radday's paper on Zechariah is strongly criticised by Portnoy and Petersen (1984) for interpreting statistically significant differences between the values of certain measures of linguistic behaviour as differences between authors. Portnoy and Petersen maintain that Radday has made some major statistical errors which in effect invalidate the firm conclusions offered and suggest that it was possible to propose statistical models, in particular multivariate techniques, which were more appropriate to the task of analyzing the book of Zechariah. They conclude that, despite the problem of small text size, both statistical analysis and tradition point to three distinct units in the book: Chapters 1–8, 9–11 and 12–14.

Radday's paper on the Book of Judges drew on 38 distinct stylistic features and concluded that statistical analysis to support the views of biblical scholars concerning the composition of the book, namely that its Main Body (Chapters 3–12) and the Samson Cycle (Chapters 13–16) appear to be by different hands.

The work done by Radday and his colleagues (1985) on the unity of Genesis has caused much comment. The book of Genesis contains three kinds of discourse: narration, direct discourse of the deity and direct discourse by humans. The material can also be cross-classified into three levels: pre-history (E), patriarchal (P) and Joseph Stories (J), and any variation due to author must be separated out from that due to other factors. Radday and his collaborators adopt two approaches in their analysis. One is to take hypotheses suggested by biblical scholarship and then to test them statistically. The second approach is to search for structure in the text without benefit of prior hypotheses. Here the authors divide Genesis into 96 segments each of about 200 words and compute about 40 stylistic variables for each segment. Multivariate methods are applied to the resulting data matrix and Radday's team concludes that Genesis was probably a unity as far as authorship was concerned, with no clear groupings emerging.

Portnoy (1988) notes the strong agreement by almost all biblical scholars that there are multiple sources in Genesis and argues that single authorship should not be accepted on the basis of statistical evidence alone. He queries the application of statistics to biblical Hebrew, considering that linguistic measurements often either fail to indicate accepted differences or indicate differences between sections accepted as singly authored. Portnoy concludes:

> Given the enormous diversity and limited quantity of biblical writings it is difficult to imagine how any measurements can be sufficiently calibrated to distinguish authorship. Until much more is known about the statistical linguistics of biblical Hebrew, statistical analysis of biblical material should be used with extreme care and discretion.

Weitzman (1988) is also concerned about the fact that Radday was apparently attempting to prove the unity of Genesis by the absence of stylistic differences between portions which Bible critics attribute to the three supposed sources J, E and P. Weitzman points out that Radday and his group did find statistically significant differences but that they then tried to explain this variation away with reference to the fact that it was no greater than the variation between human speech, divine speech and narrative which could not be attributed to separate authors. Weitzman concludes:

> Human speech, divine speech and narrative, which represent three separate genres, differ considerably, and if Radday observes comparable differences between the J, E and P portions, which do not differ in genre, this suggests – but does not prove – multiple authorship.

Bee (1971, 1972) has suggested using statistics based on verb counts to analyze whether Hebrew passages derive from written or oral material. If x_i is the word count up to the occurrence of the i'th verb (i = 1, 2 . . . n) for a passage containing n verbs, then Bee plots i against x_i, fits a straight line to the data obtained, and concerns himself with S, the sum of squared deviations of the plotted points from the fitted straight line. Bee argues that the verb rate is a component of literary style which conveys the sense of action of a passage. He finds that in many passages of considerable length, the verb rate is effectively constant for a given author. As well as counting words up to the i'th verb, Bee also counts syllables, both stressed and unstressed, and applies this method of analysis to the Masoretic Text of the Old Testament. Determining the stressing of a text is a rather subjective process, however, although Bee argues that the stresses counted should be those employed in scansion.

Bee proceeds on the assumption that the data type which gives the least value of S characterizes the author's method of composition and style, the composition being called "ictal" where stress counts up to the i'th verb provide the lowest S value and "verbal" when word counts provide the lowest S value. Bee provides a somewhat rudimentary significance test to decide whether or not one set of counts provide a significantly lower S value than the other. He found from a study of Old Testament prophets that most of the material was of ictal composition and he also detected in many Old Testament books evidence of changes in preacher from abrupt changes in slope of the verb rate plots. Weitzman (1981) notes, however, that Bee's conclusions suffer from inadequate sampling theory, the use of critical values determined in a very ad hoc manner and lack of evidence that the measures reasonably reflect source differences.

Many criticisms of authorship studies such as those discussed above have centred on the use of random sampling theory as a basis for significance tests. We rarely, if ever, have a text which can be regarded as a random sample of all that the author has written, most works may instead be regarded as a segment from a time series since writings are produced in a more or less sequential fashion. This point is taken up by Bartholomew (1988):

> A more promising line is to see the production of a literary work as a stochastic process constrained by the grammar and syntax of the language and by the customs of the literary genre. A text would then be characterized by the parameters of the process and the stylometric value of the analysis would depend on there being a few simple invariants of literary style which would characterize the author's style.

Bartholomew comments that the Sichel distribution, introduced earlier in this paper, arises from a stochastic distribution and may prove fruitful.

Pollatschek and Radday (1985) used the Sichel distribution in an analysis of the text of the Book of Genesis. The vocabulary of this Book may be separated into the three categories briefly mentioned earlier, namely the words of the narrator (N), human direct speech (H) and Divine direct speech (D). On an α–θ plot involving several textual samples from all three categories, Pollatschek and Radday found that the N, H and D samples were consistent within themselves yet the three categories were unmistakably separated and exhibited three entirely different ways of behaviour as to their vocabulary properties. The authors concluded that this was an almost inexplicable literary feat which would have been

difficult to achieve by one writer, let alone a Documentary School of several writers who some claim have had a share in composing Genesis. Other applications of the Sichel model in the original Hebrew include the Books of Lamentations and Zechariah.

19. Conclusion

Because authors are influenced by subject matter and because their powers develop with maturity and experience, attribution methods are likely to be most reliable when the texts of known authorship are of the same date and genre as the anonymous work. Smith (1990) lists some general principles which should be observed in authorship studies:

1) The onus of proof lies entirely with the person making the ascription.
2) The argument for adding something to an author's canon has to be vastly more stringent than for keeping it there.
3) If doubt persists, an anonymous work must remain anonymous.
4) Avoidance of a false attribution is far more important than failing to recognise a correct one.
5) Only works of known authorship are suitable as a basis for attributing a disputed work.
6) There are no short-cuts in attribution studies.

All authorship studies begin with a choice of criteria believed to characterize authors. One should probably not believe that any single set of variables is guaranteed to work for every problem, so researchers must be familiar with variables that have worked in previous studies as well as the statistical methods to determine their effectiveness for the current problem.

On a personal note, I believe that the future for stylometry over the next decade will lie in the development of connectionist approaches entailing the extensive use of Artificial Neutral Networks. Authorship attribution can be regarded as a special kind of pattern recognition where the pattern that is being searched for is the specific feature of the text that is thought to distinguish one author from others. The connectionist approach concentrates on learning the sequences of features that appear in text. If an artificial neural network is given as much information about a text as possible in the form of attributes such as punctuation, phrases, parts of speech, etc, when it is being trained, then the rules or tests that it uses to distinguish between authors are very likely to be the best methods for determining authorship. It is hoped that these rules will be general ones in the sense of being independent of genre, subject and time period.

Stylometry presents no threat to traditional scholarship. In the context of authorship attribution, stylometric evidence must be weighed in the balance along with that provided by more conventional studies. Stylometry does, all the same, have a role to play, despite the suspicions of those who mistrust the application of statistical and computing techniques to literature and the analysis of texts. It seems appropriate to conclude with Burrows' timely reminder (1992) that:

> Literary theorists . . . are not entitled to deny that literary works are marked by the particular stylistic habits and, by a not unreasonable inference, the intellectual propensities of their authors.

References

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

Bailey, R.W. "Authorship Attribution in a Forensic Setting." *Advances in Computer-aided Literary and Linguistic Research*. Eds. D.E. Ager, F.E. Knowles and J. Smith. Birmingham: AMLC, 1979.

Baker, J.C. Pace. "A Test of Authorship Based on the Rate at Which New Words Enter an Author's Text." *Journal of the Association for Literary and Linguistic Computing*, 3, 1 (1988), 36–39.

Bartholomew, D.J. "Probability, Statistics and Theology." *Journal of the Royal Statistical Society*, A, 151, 1 (1988), 137–78.

Bee, R.E. "Statistical Methods in the Study of the Masoretic Text of the Old Testament." *Journal of the Royal Statistical Society*, A, 134, 4 (1971), 611–622.

Bee, R.E. "A Statistical Study of the Sinai Periscope." *Journal of the Royal Statistical Society*, A, 135, 3 (1972), 406–421.

Bender, T.K. and S.M. Briggum. "Quantitative Stylistic Analysis of Impressionist Style in Joseph Conrad and Ford Maddox Ford." In *Computing in the Humanities*. Ed. R.W. Bailey. North-Holland, 1982.

Bennett, P.E. "The Statistical Measurement of a Stylistic Trait in *Julius Caesar* and *As You Like It*." In *Statistics and Style*. Eds. L. Dolezel and R.W. Bailey. New York: American Elsevier, 1969.

Boreland, H. and P. Galloway. "Authorship, Discrimination and Clustering: Timoneda, Montesino and Two Anonymous Poems." *Association for Literary and Linguistic Computing Bulletin*, 8 (1980), 125–151.

Brainerd, B. "On the Distinction Between a Novel and a

Romance: A Discriminant Analysis." *Computers and the Humanities*, 7 (1973), 259–270.

Brainerd, B. *Weighing Evidence in Language and Literature: A Statistical Approach*. University of Toronto Press, 1974.

Brainerd, B. "Two Models for the Type-Token Relation with Time Dependant Vocabulary Reservoir." In *Vocabulary Structure and Lexical Richness*. Eds. P. Thoiron, D. Serant and D. Labbe. Paris: Champion-Slatkine, 1988.

Brinegar, C.S. "Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship." *Journal of the American Statistical Association*, 58 (1963), 85–96.

Bruno, A.M. *Toward a Quantitative Methodology for Stylistic Analyses*. University of California Press, 1974.

Burrows, J.F. "Word Patterns and Story Shapes: The Statistical Analysis of Narrative Style." *Journal of the Association for Literary and Linguistic Computing*, 2, 2 (1987), 61–70.

Burrows, J.F. and A.J. Hassall. "Anna Boleyn and the Authenticity of Fielding's Feminine Narratives." *Eighteenth Century Studies*, 21 (1988), 427–453.

Burrows, J.F. "Computers and the Study of Literature." In *Computers and Written Texts*. Ed. C.S. Butler. Oxford: Blackwell, 1992.

Cox, D.R. and L. Brandwood. "On a Discriminating Problem Connected with the Works of Plato." *Journal of the Royal Statistical Society*, B, 21 (1959), 195–200.

Damerau, F.J. "The Use of Function Word Frequencies as Indicators of Style." *Computers and the Humanities*, 9 (1975), 271–280.

Delcourt, C. "On Vocabulary Curves." *Association for Literary and Linguistic Computing Journal*, 2 (1981), 13–24.

Ellegard, A. *A Statistical Method for Determining Authorship: The Junius Letters, 1769–1772*. Gothenburg: University of Gothenburg, 1962.

Fucks, W. "On the Mathematical Analysis of Style." *Biometrika*, 39 (1952), 122–129.

Fucks, W. and J. Lauter. "Mathematische Analyse des Literarischen Stils." In *Mathematik und Dichtung*. Eds. H. Kreuzer and R. Gunzenhausers. Munich: Nymphenburger Verlagsbuckhandlung, 1965.

Grayston, K. and G. Herdan. "The Authorship of the Pastorals in the Light of Statistical Linguistics." *New Testament Studies*, 6 (1959), 1–15.

Gregory, M.J. "An Approach to the Study of Style." *Linguistics and Style*. Eds. N. Enkvist, J. Spencer and M.J. Gregory. University of Oxford Press, 1964.

Herdan, G. "A New Derivation and Interpretation of Yule's 'Characteristic' K." *Journal of Applied Mathematics and Physics*, 6 (1955), 332–334.

Herdan, G. *Quantitative Linguistics*. London: Butterworths, 1964.

Herdan, G. *The Advanced Theory of Language as Choice and Chance*. New York: Springer-Verlag, 1966.

Holmes, D.I. "Vocabulary Richness and the Prophetic Voice." *Literary and Linguistic Computing*, 6, 4 (1991), 259–268.

Holmes, D.I. "A Stylometric Analysis of Mormon Scripture and Related Texts." *Journal of the Royal Statistical Society (A)*, 155, 1 (1992), 91–120.

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

Hubert, P. and D. Labbe, D. "A Model of Vocabulary Partition." *Journal of the Association for Literary and Linguistic Computing*, 3, 4 (1988), 223–225.

Johnson, R. "Measures of Vocabulary Diversity." In *Advances in Computer-aided Literary and Linguistic Research*. Eds. D.E. Ager, F.E. Knowles and M.W.A. Smith. Birmingham: AMLC, 1979.

Kemp, K.W. "Aspects of the Statistical Analysis and Effective Use of Linguistic Data." *Association for Literary and Linguistic Computing Bulletin*, 4 (1976), 14–22.

Kenny, A. *A Stylometric Study of the New Testament*. Oxford University Press, 1986.

Kjetssa, G. "*And Quiet Flows the Don* Through the Computer." *Association for Literary and Linguistic Computing Bulletin*, 7 (1979), 248–256.

Kjetssa, G. "Written by Dostoyevsky." *Association for Literary and Linguistic Computing Journal*, 2 (1981), 25–33.

Ledger, G.R. *Re-counting Plato: A Computer Analysis of Plato's Style*. Oxford: Clarendon, 1989.

Mandelbrot, B. "A Note on a Class of Skew Distribution Functions: Analysis and Critique of a Paper by H.A. Simon." *Information and Control*, 2 (1959), 90–99.

Mendenhall, T.C. "The Characteristic Curves of Composition." *Science*, IX (1887), 237–249.

Miles, J. and H. C. Selvin. "A Factor Analysis of the Vocabulary of Poetry in the Seventeenth Century." In *The Computer and Literary Style*. Ed. J. Leed. Ohio: Kent State University Press, 1966.

Morton, A.Q. "The Authorship of Greek Prose." *Journal of the Royal Statistical Society*, A, 128 (1965), 169–233.

Morton, A.Q. *Literary Detection*. New York: Scribners, 1978.

Morton, A.Q. "Once. A Test of Authorship Based on Words which are not Repeated in the Sample." *Journal of the Association for Literary and Linguistic Computing*, 1, 1 (1986), 1–8.

Morton, A.Q. and J. McLeman. *The Genesis of John*. Edinburgh: St Andrew's Press, 1980.

Mosteller, F. and D.L. Wallace. "Inference and Disputed Authorship: The *Federalist*." Reading, MA: Addison-Wesley, 1964.

Muller, C. "Calcul des Probabilités et Calcul d'un Vocabulaire." *Travaux de Linguistique et de Littérature* (1964), 235–244.

Muller, C. "Lexical Distribution Reconsidered: the Waring-Herdan Formula." In *Statistics and Style*. Eds. L. Dolezel and R.W. Bailey, New York: American Elsevier, 1969.

Muller, C. "Peut-on estimer l'étendue d'un lexique?" *Cahiers de Lexicologie*, 27 (1975), 3–29.

Oakman, R.L. *Computer Methods for Literary Research*. Columbia: University of South Carolina Press, 1980.

Pollatschek, M. and Y.T. Radday. "Vocabulary Richness and Concentration in Hebrew Biblical Literature." *Association for Literary and Linguistic Computing Bulletin*, 8 (1981), 217–231.

Pollatschek, M. and Y.T. Radday. "Vocabulary Richness and Concentration." In *Genesis: An Authorship Study*. Eds.

Y.T. Radday and H. Shore. Rome: Biblical Institute Press, 1985.

Portnoy, S. "Reply to Professor Bartholomew." *Journal of the Royal Statistical Society*, A, 151, 1 (1988), 172.

Portnoy, S. and D.L.Petersen. "Biblical Texts and Statistical analysis: Zechariah and Beyond." *Journal of Biblical Literature*, 103 (1984), 11–21.

Radday, Y.T. *The Unity of Isaiah in the Light of Statistical Linguistics*. Gerstenberg: Hindlesheim, 1973.

Radday, Y.T. and D. Wickmann. "The Unity of Zechariah in the Light of Statistical Linguistics." *Zeit Alttestamentliche Wissenschaft*, 87 (1975), 30–55.

Radday, Y.T. and M. Pollatschek. "Frequency Profiles: A Key to the M. Pollatschek Structure of Lamentations." *Balsanut Hofsit*, 12 (1977), 24–35.

Radday, Y.T., D. Wickmann, G. Leb, and S. Talman. "The Book of Judges Examined by Statistical Linguistics." *Biblica*, 58 (1977), 469–499.

Radday, Y.T. and H. Shore. *Genesis: An Authorship Study in Computer-assisted Statistical Linguistics*. Rome: Biblical Institute Press, 1985.

Ratkowsky, D.A. and L. Hantrais. "Tables for Comparing the Richness and Structure of Vocabulary in Texts of Different Lengths." *Computers and the Humanities*, 9 (1975), 69–75.

Sichel, H.S. " On a Distribution Representing Sentence-Length in Written Prose." *Journal of the Royal Statistical Society (A)*, 137 (1974), 25–34.

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

Sichel, H.S. "Word Frequency Distributions and Type-Token Characteristics." *Mathematical Scientist*, 11 (1986), 45–72.

Simpson, E.H. "Measurement of Diversity." *Nature*, 163 (1949), 688.

Smith, M.W.A. "Recent Experience and New Developments of Methods for the Determination of Authorship." *Association for Literary and Linguistic Computing Bulletin*, 11 (1983), 73–82.

Smith, M.W.A. "An Investigation of the Basis of Morton's Method for the Determination of Authorship." *Style*, 19, 3 (1985a), 341–368.

Smith, M.W.A. "An Investigation of Morton's Method to Distinguish Elizabethan Playwrights." *Computers and the Humanities*, 19, 1 (1985b), 3–21.

Smith, M.W.A. "Hapax Legomena in Prescribed Positions: An Investigation of Recent Proposals to Resolve Problems of Authorship." *Journal of the Association for Literary and Linguistic Computing*, 2, 3 (1987a), 145–152.

Smith, M.W.A. "The Authorship of Pericles: New Evidence for Wilkins." *Journal of the Association for Literary and Linguistic Computing*, 2, 4 (1987b), 221–30.

Smith, M.W.A. "Attribution by Statistics: A Critique of Four Recent Studies." *Revue, Informatique et Statistique dans les Sciences Humaines*, 26 (1990), 233–251.

Smith, M.W.A. "The Authorship of *The Raigne of King Edward the Third*." *Literary and Linguistic Computing*, 6, 3 (1991a), 166–174.

Smith, M.W.A. "The Authorship of *The Revenger's Tragedy*." *Notes and Queries*, 38, 4 (1991 b), 508–513.

Somers, H.H. "Statistical Methods in Literary Analysis." In *The Computer and Literary Style*. Ed. J. Leed, Ohio: Kent State University Press, 1966.

Tallentire, D.R. *An Appraisal of Methods and Models in Computational Stylistics, with Particular Reference to Author Attribution*. PhD thesis. University of Cambridge, 1972.

Tallentire, D.R. "Towards an Archive of Lexical Norms – A Proposal." In *The Computer and Literary Studies*. Eds. A.J. Aitken, R.W. Bailey and N. Hamilton-Smith. Edinburgh University Press, 1973.

Tallentire, D.R. "Confirming Intuitions about Style Using Concordances." In *The Computer in Literary and Linguistic Studies*. Eds. A. Jones and R.F. Churchouse. University of Wales Press, 1976.

Thoiron, P. "Diversity Index and Entropy as Measures of Lexical Richness." *Computers and the Humanities*, 20, 3 (1986), 197–202.

Ule, L. "Recent Progress in Computer Methods of Authorship Determination." *Association for Literary and Linguistic Computing Bulletin*, 10 (1982), 73–89.

Wake, W.C. "Sentence-Length Distributions of Greek Authors." *Journal of the Royal Statistical Society*, A, 120 (1957), 331–346.

Weitzman, M.P. "Reply to Professor Bartholomew." *Journal of the Royal Statistical Society*, A, 151, 1 (1988), 173.

Williams, C.B. "A Note on the Statistical Analysis of Sentence-Length as a Criterion of Literary Style." *Biometrika*, 31 (1940), 356–361.

Williams, C.B. *Style and Vocabulary: Numerical Studies*. Griffin, 1970.

Yule, G.U. "On Sentence-Length as a Statistical Characteristic of Style in Prose, with Application to Two Cases of Disputed Authorship." *Biometrika*, 30 (1938), 363–390.

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

Zipf, G.K. *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge, MA: Harvard University Press, 1932.