Advanced text authorship detection methods and their application to biblical texts

Tālis J. Putniņš[a], Domenic J. Signoriello[a], Samant Jain[a], Matthew J. Berryman[a] and Derek Abbott[a],

[a]Centre for Biomedical Engineering and
School of Electrical & Electronic Engineering,
The University of Adelaide, SA 5005, Australia.

ABSTRACT

Authorship attribution has a range of applications in a growing number of fields such as forensic evidence, plagiarism detection, email filtering, and web information management. In this study, three attribution techniques are extended, tested on a corpus of English texts, and applied to a book in the New Testament of disputed authorship. The word recurrence interval based method compares standard deviations of the number of words between successive occurrences of a keyword both graphically and with chi-squared tests. The trigram Markov method compares the probabilities of the occurrence of words conditional on the preceding two words to determine the similarity between texts. The third method extracts stylometric measures such as the frequency of occurrence of function words and from these constructs text classification models using multiple discriminant analysis. The effectiveness of these techniques is compared. The accuracy of the results obtained by some of these extended methods is higher than many of the current state of the art approaches. Statistical evidence is presented about the authorship of the selected book from the New Testament.

Keywords: Authorship attribution; stylometry; word recurrence interval; trigram Markov model; multiple discriminant analysis.

1. INTRODUCTION

Authorship attribution is the problem of identifying the author of an anonymous text, or text whose authorship is in doubt[1]. The techniques of authorship attribution have a range of applications in a growing number of fields such as forensic evidence, plagiarism detection, email filtering, and solving literary debates where authorship is disputed. Due to the vast repositories of electronic text that have become available on the Internet recently, an application of growing interest is web information management. In this field, authorship attribution techniques are beginning to play a role in areas such as information retrieval, information extraction, categorizing documents in large repositories by their author, and question answering[2]. Despite the range of techniques developed, there is little consensus as to which is the most effective and a higher level of attribution accuracy is required for many of the aforementioned applications. The widespread use of electronic media means there is a need for accurate computerized techniques.

Data mining is an information extraction activity that uses a combination of statistical analysis, machine learning, artificial intelligence and modelling techniques to infer rules that allow the prediction of future results. Artificial intelligence attempts to apply human-like thought processing to statistical problems. Machine learning is a combination of artificial intelligence and statistics that allows computer programs to infer relationships from data.

Stylometry attempts to define the features of an author's style and to determine statistical methods to measure these features so that the similarity between two or more pieces of text can be analysed[3]. Authors have both conscious and subconscious aspects to their writing. Nearly all experts of literary style postulate that style is dictated by the subconscious aspects, which are more consistent and measurable, and form a unique fingerprint of a writer's work.

Combining these ideas about measurable aspects of style and data mining techniques, the approach to authorship attribution taken in this study involves firstly measuring and quantifying the subconscious aspects of writing style, then applying statistical and data mining techniques to these measures in order to detect authorship. A number of extensions,

Complex Systems, edited by Axel Bender, Proc. of SPIE
Vol. 6039, 60390J, (2006) · 0277-786X/06/$15 · doi: 10.1117/12.639281

previously unexplored, are made to existing techniques. The accuracy of the results obtained by some of these extended computerized techniques is higher than many of the current state of the art approaches. Significant statistical evidence is presented, here, on the traditionally debated authorship of *The Letter to the Hebrews*.

2. WORD RECURRENCE INTERVAL BASED METHOD

2.1. Extraction of word recurrence interval measures

2.1.1. Description

Word recurrence interval (WRI) is the term used to represent the number of words between successive occurrences of a keyword. For example, in the phrase "the cat sat on the mat" the word recurrence interval for the keyword "the" is three as there are three words between the two occurrences of the word "the". For the purpose of this study, a keyword is defined as a word that appears in a text more than five times. Upon extraction of a set of keyword WRIs, $\{x_1,.....,x_k\}$, the scaled standard deviation of these is computed using the following formula.

$$\hat{\sigma} = \frac{1}{\bar{x}}\sqrt{\sum_{i=1}^{k}\frac{(x_i-\bar{x})^2}{1-k}}\,.$$

The motivation for using the scaled standard deviation is to eliminate the dependency on word frequency when characterising word distributions thus capturing a characteristic of authorial style different to the frequency of word usage that was previously examined. This process is repeated for all keywords within a text, resulting in a set of scaled standard deviations, $\{\hat{\sigma}_1,.....\hat{\sigma}_k\}$. The $\hat{\sigma}_n$ are ranked in order of magnitude and then plotted against $\log_{10}$(rank). Hence the $\hat{\sigma}_n$ values that are compared graphically are not necessarily for the same underlying keyword. This technique was introduced by Ortuno *et al.*[15].

2.1.2. Results and conclusions

This technique was applied to a variety of English texts of known authorship to examine its effectiveness in capturing style characteristics. Two of the resulting plots are shown below.

Figure 1a shows the similarity between the works of Oscar Wilde (*The Picture of Dorian Gray and Lord Arthur Savile's Crime and Other Stories*) and Lewis Carroll (*The Adventures of Alice in Wonderland* and *Through the Looking Glass*). The plot suggests that WRI may have captured some of the style differences between the works of Carroll and Wilde. It appears that the works of Carroll have a generally greater slope and are coincident with each other for a greater length of the plot than the works of Wilde. However, this is a rather subjective analysis and the discrimination between the authors is not particularly obvious or convincing.

Other tests, using different texts or authors, contradicted the conclusion that WRI may have captured some of the style differences between the authors. An example is given in Figure 1b. In this plot the texts by Charles Dickens (*A Christmas Carol* and *Oliver Twist*) are quite dissimilar as are the works of Thomas Hardy (*Noble Dames* and *The Romantic Adventures of Milk Maid*). For most of the length of the plot, it seems as though *Dickens' A Christmas Carol* and *Hardy's the Romantic Adventures of Milk Maid* are nearly coincident. This suggests that either the standard deviation of WRI is somewhat inconsistent in characterizing style and attributing authorship or that stylistically confusing texts have been encountered.

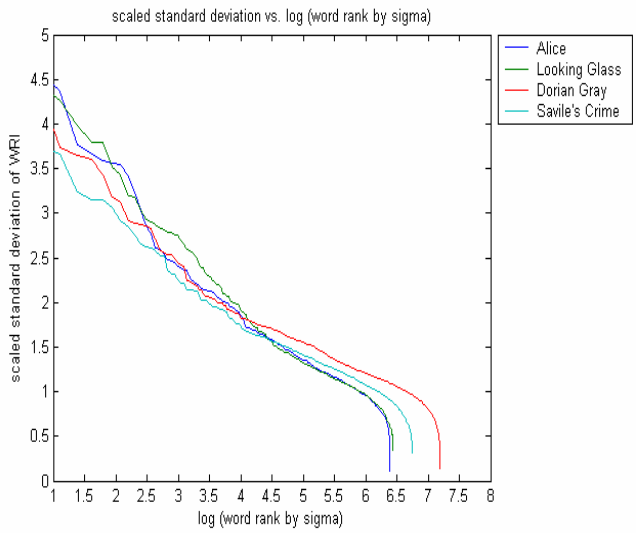

Figure 1a: Scaled standard deviation of WRI (x-axis) vs. $\log_{10}$(rank) (y-axis) for selected works of Oscar Wilde and Lewis Carroll.

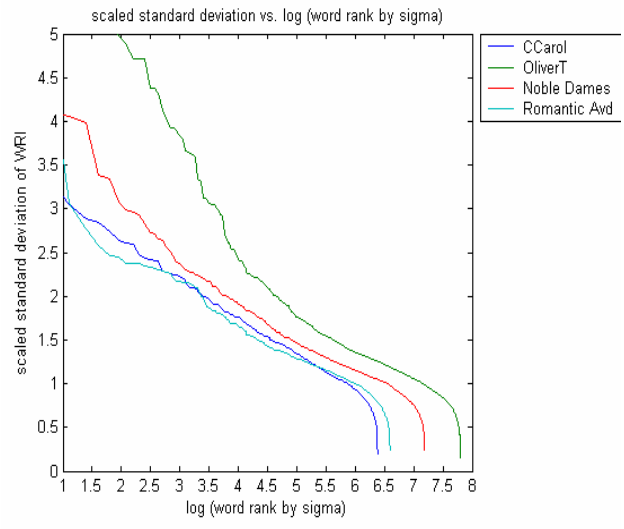

Figure 1b: Scaled standard deviation of WRI (y-axis) vs. $\log_{10}$(rank) (x-axis) for selected works of Charles Dickens and Thomas Hardy.

The conclusions reached are that the plots of standard deviation of WRI are affected by the length of the text which should be controlled for and that these plots tend to be inconsistent indicators of authorial style under subjective visual analysis. This motivates a quantitative analysis of the WRI measures, which is undertaken in the following section.

2.2. Chi-squared analysis of WRI measures

2.2.1. Description

Using the previously described method a set of scaled standard deviations of WRIs is generated for words common to all texts in the selection. These words are ranked in descending order according to the maximum of the product $f\hat{\sigma}$ for a given common keyword, where f represents the frequency or the number of times the word occurs in a text. A selection of these words is then taken with a high standard deviation and statistical significance. Having obtained the keywords, sets of variances of the keyword WRIs, $\{\hat{\sigma}^2_{11},.....,\hat{\sigma}^2_{N1}\},.....,\{\hat{\sigma}^2_{1W},.....,\hat{\sigma}^2_{NW}\}$, are generated for each text (for keywords $n = 1,.....,N$ and texts $w = 1,.....,W$). The value of χ^2 is then calculated for pairs of texts, $(l, m) \in \{1,.....,W\}\times\{1,.....W\}$ using the formula:

$$\chi^2_{lm} = \frac{1}{N_l N_m}\sum_{n=1}^{N}\frac{\left(N_m\hat{\sigma}^2_{nl} - N_l\hat{\sigma}^2_{nm}\right)^2}{\hat{\sigma}^2_{nl}+\hat{\sigma}^2_{nm}}, \text{ where } N_l = \sum_{n=1}^{N}\hat{\sigma}^2_{nl} \text{ and } N_m = \sum_{n=1}^{N}\hat{\sigma}^2_{nm}.$$

Ideally around 30 common keywords would be used. For the shorter biblical texts there are seldom that many common keywords with both a high scaled standard deviation and statistical significance. This problem is overcome by normalizing the values of Chi-squared (χ^2) by a factor of 30/N, where N represents the number of keywords. Low values of χ^2 indicate similarity between word distributions in texts.

2.2.2. Results and conclusions

Chi-squared tests were performed on English texts of known authorship by the authors Wells (*War of the Worlds* and *The Time Machine*) and Doyle (*Through the Magic Door* and *The Hounds of Baskerville*). The results are shown in the table below.

Table 1: Chi-squared values for pairs of selected texts by H.G.Wells and Doyle.

	War of the Worlds	The Time Machine	Magic Door	Baskerville
War of the Worlds	0	4.83	7.74	5.22
The Time Machine	4.83	0	5.58	4.64
Magic Door	7.74	5.58	0	0.845
Baskerville	5.22	4.64	0.845	0

The χ^2 value between the two texts written by Doyle, 0.845, is by far the lowest value in the table for any pairs of texts indicating the greatest amount of similarity. In this case sense the χ^2 tests have successfully identified the similarity between the texts written by Doyle. However, the χ^2 value between the two texts written by Wells, 4.83, is significantly higher—in fact higher than a pair of texts of different authorship (*The Hounds of Baskerville* and *The Time Machine*). This suggests that the χ^2 tests have not successfully identified similarity between the texts written by Wells —thus this shows the same inconsistency that was identified in plots of standard deviation of WRI.

3. TRIGRAM MARKOV MODEL METHOD

Markov chains are used in a variety of areas in mathematics and engineering. They have also been widely used in linguistics and speech processing[12]. In the fields of linguistics and speech processing, Markov chains have been used mainly in stochastic text generation, which can be applied to spelling correction, handwriting recognition and most recently mobile phone texting[13]. Previous studies have shown Markov chains are a useful tool in determining text authorship[12, 14].

3.1. Description of method

A Markov model assigns probabilities to all sequences of words using the memoryless assumption that each state in a process is dependent only on the previous state. Third-order models, known as trigram models, are based on the assumption that word occurrence depends only on the immediately preceding two words. Defining processes by state transitions, the memoryless property simplifies calculations, thus increasing the ability to analyze such processes.

Previous studies have encoded states as single characters or letters[12, 14]. Such a model assumes that each letter or character in a text is dependent on the letter or character preceding it. Extending this idea by using the fact that lexical or token level measures are effective in characterizing authorial style, words were chosen as the states of this model. Trigram model states (trigrams) are encoded as pairs of words. For a vocabulary of two words—'a' & 'b' there are four states 'aa', 'ab', 'ba' and 'bb'. The probability of the transition from aa to ab is P(ab | aa) = P(b | aa). With a vocabulary of n words, there are 2^n possible states. The exponential nature of the state space has serious implications on the ability to process transition probability measures.

A key decision is how to use the state transition probabilities to determine text authorship. The entropy method used by Khmelev[12] was chosen as it has previously been applied to text authorship detection and is relatively simple. It is an approximation of the exact regeneration probabilities and calculated as follows.

Let p_{kl}^i be the probability of a transition from state k to state l in a text by author i, and

Let Q_{kl}^x be the number of transitions from state k to state l in the text of unknown authorship.

Suppose the author of text x could be either of the authors $0,1 \ldots n$.

Calculate $\Lambda_i = -\sum_{\forall k,l} Q_{kl}^x \ln(p_{kl}^i)$ for all $i=0,1 \ldots n$.

The best estimate of the author of text x is author X, where $X = \arg\max_{i=0,1\ldots n}(\Lambda_i)$.

3.2. Results and conclusions

In order to evaluate the effectiveness of this technique the same large corpus of English fictional texts was used. For convenience the authors are labeled A, B, C, H, R and Z corresponding to Sir Arthur Conan Doyle, B. M. Bower, Charles Dickens, Henry James, Richard Harding Davis and Zane Grey respectively.

Three different tests were performed on these data. The first test included punctuation and words with letters of different cases. For the second test, all punctuation was removed from the texts and all letters were converted to lower case prior to classification. For the third test all punctuation and any words that were not entirely in lower case were removed from the texts prior to classification. These tests allow comparison with Khmelev[14], who used similar tests with single letter states. If a text was correctly classified it was given a rank of one, if the true author was predicted as the second most likely candidate the text was given a rank of two and so on down to the worst case misclassification scoring a rank of six. The closer the ranking of a text is to one, the more accurate the authorship attribution technique has been.

Initially one text from each author was selected as the texts of known authorship used for training the classification model with which the remaining texts were classified. The average ranks given to the texts classified are shown in Figure 2a.

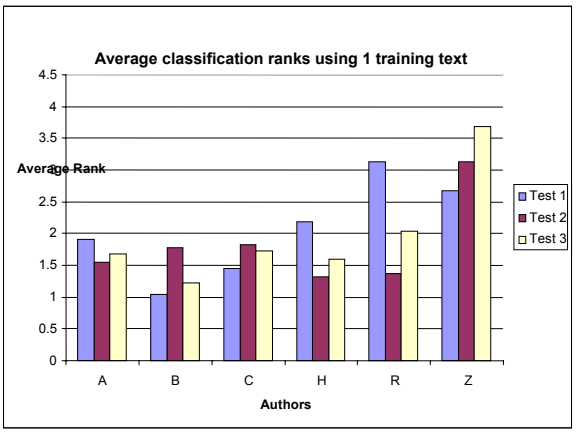

Figure 2a: Average classification rank of texts (y-axis) by each of six authors (x-axis) using a single text from each author as model training data.

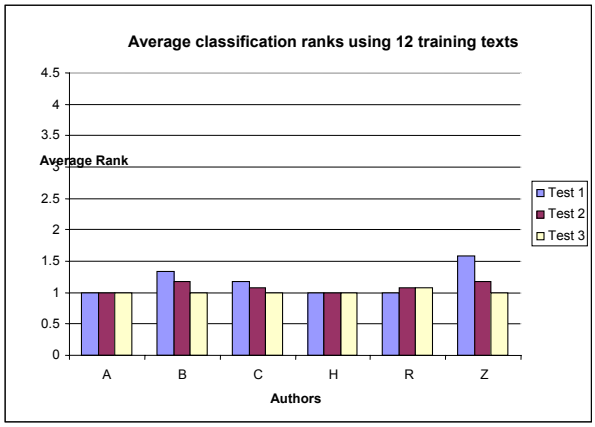

Figure 2b: Average classification rank (y-axis) of texts by each of six authors (x-axis) using 12 texts from each author as model training data.

By chance alone (assumes the classification procedure has no authorship attribution ability) there is an equal probability of a text being attributed to any of the authors. This probability is 1/6. Using the ranking system, the expected average rank, *R,* produced by chance is equal to 3.5. From the results above, even though a training set of only one text from each author has been used to construct the classification model, its accuracy is higher than chance for most tests. Overall the classification accuracy using this one text as training data was 41.0%. Using a different text as training data gave a classification accuracy of 45.1% making the average accuracy for one text training data 43.0%.

The size of the training set was increased to four and then 12 texts from each author, which were concatenated and used as the training text (control text). Two random samples were taken for each training data size and the accuracies found on each sample averaged. As an example, the resulting ranks for one of the samples of 12 texts per author training data are shown above in Figure 2b.

The average classification accuracy (correct classification on first prediction) for all tests performed using four texts per author as training data was 66.3% and using 12 texts per author, the classification accuracy was 88.3%. These results show a clear upward trend in the classification accuracy as the size of the training data is increased. This is also seen in Figure 8 above in the form of the consistently lower rankings. This effect is consistent with the findings using the first authorship attribution technique of multiple discriminant analysis of function word frequencies. It should be noted that the classification accuracy of 88.3% was achieved using 12 texts per author as training data and the classification

accuracy obtained with a greater size of training data was not explored. Hence the full potential of this technique is not yet known. The accuracy is expected to increase with a larger sized set of training data as the first technique's baseline testing indicated maximal classification accuracy is only achieved with 16 texts per author or more of training data. From these results it cannot be determined which, of the three types of test is the most effective. Different sizes of training data and authors favor different types of test.

4. MULTIPLE DISCRIMINANT ANALYSIS OF FUNCTION WORD FREQUENCIES

4.1. Description of method

4.1.1. Overview

Adopting a stylometric approach, the problem of authorship attribution can be broken down into two distinct modules: extraction of style markers and a classification procedure. In this approach the frequency of occurrence of function words and sequences of function words were used as the style markers and for the classification procedure, multiple discriminant analysis was used.

4.1.2. Style marker extraction

Style markers can be generally grouped into three levels: the lexical level[4], syntactic or grammatical level[5], and the language level[2]. The most important approaches to authorship attribution are exclusively based on lexical measures that either represent the vocabulary richness of the author or simply comprise frequencies of occurrence of function (or context-free) words[3]. Function words, for example "the", "is" and "of", have little lexical meaning or ambiguous meaning. Their usage is generally context independent and they serve to express grammatical relationships with other words in a sentence. Function word frequencies have been shown to be a reliable discriminating factor to distinguish between authors[6, 7]. In addition to the aforementioned reasons, function word frequencies were chosen to use as the style markers due to their low computational cost relative to other markers, the ability to automate their extraction and the wide applicability of the technique as a result of their genre independent nature.

In the automated function word frequency extraction every text in the corpus is equally represented regardless of text length making the extracted function words reflective of the entire corpus rather than some subset. The frequency of occurrence of sequences of function words, which can be considered a marker from the phrase or grammatical level, was also examined but found to be less effective at distinguishing between authors.

4.1.3. Multiple discriminant analysis

Multiple discriminant analysis (MDA) generates discriminant functions from pre-classified training data that can then be used to predict the group membership of unseen cases. The discriminant functions are constructed with the aim of forcing the groups to be as statistically distinct as possible by maximizing the between group variance, while minimizing the within group variance. This can be performed as a stepwise procedure where the most correlated independent variable is selected first, the variance in the dependent variable is removed, and then the independent variable which most highly correlates with the remaining variance in the dependent is selected. This continues until the addition of a variable does not increase the R-squared by a specified statistically significant amount.

There are multiple methods of actually classifying cases. One of the simplest is by using the classification functions themselves. However, in the authorship attribution technique developed in this study, a slightly more complicated method of classification is used based on Mahalanobis distance—a measure of multivariate similarity based on correlations between variables. A group centroid is defined as the mean value for the discriminant function scores for that group. Squared Mahalanobis distance of a case with observed stylometric vector $\mathbf{x}$ from the group with mean stylometric vector μ_x is defined by:

$$d^2 = (\mathbf{x} - \mu_x)^T \mathbf{C}_x^{-1} (\mathbf{x} - \mu_x),$$

where $\mathbf{C}_x$ is the covariance matrix of $\mathbf{x}$. Upon calculating the Mahalanobis distance of a case to each group, the case is classified as belonging to the group for which its Mahalanobis distance is smallest.

4.1.4. Leave-one-out cross-validation

Cross-validation is particularly important in guarding against the two serious problems that often arise in data mining—testing hypotheses suggested by the data rather than hypotheses developed independently of the data and model overfitting to the data. In the context of authorship attribution, data (texts) are often scarce, making it difficult to perform normal cross-validation due to insufficient training and test examples. This can be overcome using leave-one-out (LOO) cross-validation—where each text is in turn left out of the training set while constructing a classification model which is used to classify the left out text. The classification accuracy is then simply calculated as the ratio of correctly classified texts to total number of texts. Such a form of cross-validation has the advantage that all the data can be used for training rather than having a sub-set held back. It provides an unbiased measure the accuracy with which additional texts would be correctly classified. Hence it can be used to detect model overfitting.

4.2. Baseline testing

4.2.1. English fictional literature text corpus

In order to establish a baseline and evaluate the effectiveness of this technique a large corpus of texts was obtained that does not suffer from the deficiencies of the New Testament such as limited text length, very limited size of training data set, disputed authorship, modifications to the original text, and disputed language of original texts. This allowed the effect of introducing the deficiencies to be examined in a controlled manner. Once these effects were understood the technique could be applied to imperfect corpuses, such as the New Testament, with considerations made for the imperfections.

The full corpus of texts in English was obtained from the Project Gutenberg archives[8]. It contains a total of 156 texts—26 texts by each of the six authors: Zane Grey, Henry James, B. M. Bower, Sir Arthur Conan Doyle, Richard Harding Davis and Charles Dickens. All six authors wrote fictional literature in English around the same period in time (late 19th century to early 20th century). Hence genre, language of original texts and period of writing are kept constant removing the possibility of being able to discriminate between authors on these grounds, thus forcing the technique to discriminate between authors based on underlying style characteristics. Due to the long length of many of these books, only the first approximately 5,000 words from each book were used.

The corpus of English fictional texts was put through the style marker extraction process and then non-stepwise MDA. The number of function words used in the stylometric vector was chosen to be 65 initially, consistent with the findings of Stamatos *et al.*[5] that classification accuracy is maximized at approximately this number of function words. LOO cross-validation was performed resulting in a classification accuracy of 98.7%. In other words, each of the 156 texts in the corpus was in turn removed from the corpus, had a classification model constructed on the remaining 155 texts, then classified by the classification model and all but two texts were correctly classified. As a baseline, this classification accuracy from cross-validation is significantly higher than the range of classification accuracies reported by the majority of previous studies into authorship attribution including those using lexical style markers. This is perhaps at least partly due to experimentally very favourable text corpus of a relatively large number of relatively long texts.

4.2.2. The effects of the number of style markers and model overfitting

The number of style markers able to be extracted is virtually unlimited, but the number used in the classification procedure, particularly non-stepwise multiple discriminant analysis, is an important factor in determining classification accuracy. One cause of overfitting a classification model, causing a false model that performs poorly when classifying new cases, is having too many discriminating variables relative to the number of cases.

Both stepwise and non-stepwise MDA were applied to varying numbers of function words for training data sizes of 26 texts per author (the entire English fictional text corpus) and ten texts per author (an intentionally small random sample from the corpus). The plots of the resulting classification accuracies are shown in Figures 3a and 3b.

In both plots classification accuracy initially increases as the number of function words used increases. Using stepwise MDA the accuracy achieves a plateau at close to 100% accuracy, whereas with non-stepwise MDA the accuracy peaks

and begins to decrease at a certain point. This is explained by model overfitting caused by the increasing ratio of variables to the number of cases.

The reason stepwise MDA does not suffer from overfitting as non-stepwise MDA does is that it only adds variables to the classification model if they make a statistically significant contribution to minimising the specified measure of error (such as inter-group Mahalanobis distance from the group centroid) or maximising intra-group variability. Stepwise MDA will stop adding variables to the classification model after a critical point defined by the level of statistical significance is reached, thus preventing overfitting.

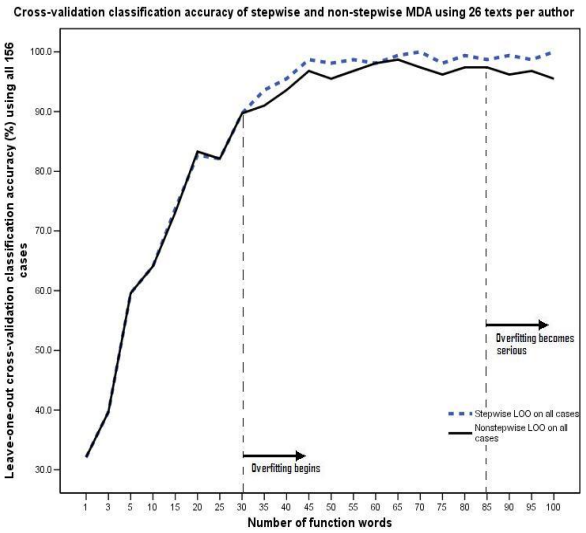

Figure 3a: LOO cross-validation classification accuracy (y-axis) of both stepwise and non-stepwise MDA using 26 texts per author as the training data set (the entire corpus) varying the number of function words used (x-axis).

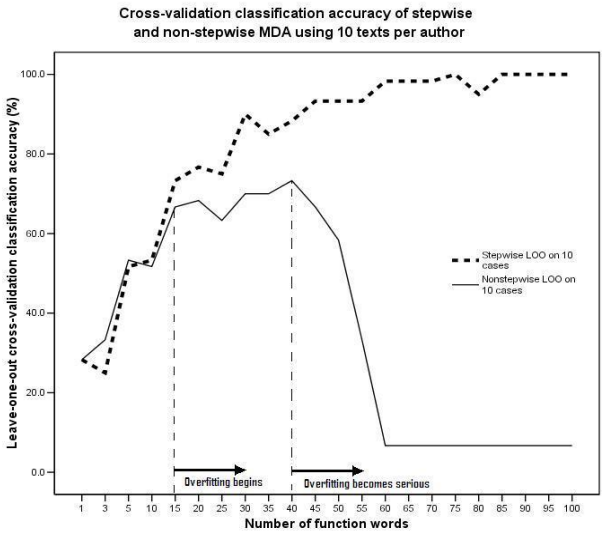

Figure 3b: LOO cross-validation classification accuracy (y-axis) of both stepwise and non-stepwise MDA using ten texts per author as the training data set (randomly sampled from corpus) varying the number of function words used (x-axis).

These findings of model overfitting are consistent with those of Stamatos *et al.*[5] who concluded classification accuracy is maximized at 60 function words (in their study, an accuracy of 77%). However, by presenting the two different tests differing only by the amounts of training data used, it can be seen that it is not the number of variables used alone that affects accuracy, but rather the ratio of variables to training cases. The second plot, having used less training data, shows far more serious overfitting effects that begin to take place at smaller numbers of function words. It was found that the point at which overfitting becomes serious in non-stepwise MDA occurs at an independent variable to cases per author ratio of approximately four. In other words, when using non-stepwise MDA classification accuracy is maximised using four times more function words than the number of texts per author in the training data.

4.2.3. The effect of the training set size and text length

Of critical importance in determining the applicability of techniques developed for authorship attribution to real problems is the size of the training set and the text length as these are often limited in practice. Of particular interest is the question of what is the minimum number of texts required as samples of an author's work and the minimum length of these texts for authorship attribution to achieve acceptable levels of classification accuracy.

In the first test all experimental variables were held constant except for the number of texts per author (the training data set size) which was varied by taking random samples of different numbers of texts by each author. In the second test all experimental variables were held constant except for the length of the texts. The results are shown in Figures 4a and 4b.

In the first test (Figure 4a), two measures of classification accuracy were obtained. Firstly the selected cases were themselves used to obtain a measure of the LOO cross-validation classification accuracy (solid line). Secondly, the unselected cases were classified using the classification model constructed using the selected cases (dashed line). The

most reliable and stable measures of cross-validation accuracy are obtained when maximising the number of cases being classified. When using less than half of the texts in the corpus for training the classification model (less than 13 texts per author) the unselected text cross-validation measure is the best measure of accuracy and when using more than half of the texts in the corpus for training the classification model (greater than 13 texts per author) the selected text LOO cross-validation measure is the best measure of accuracy. By this argument, the best measure of classification accuracy is shown on the plots as the highlighted portion of the line.

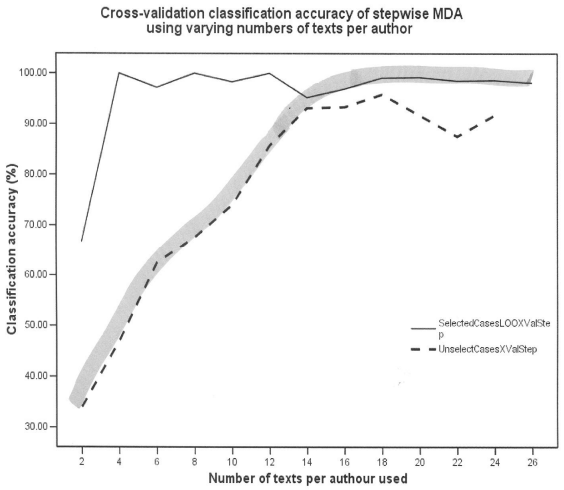

Figure 2: The effect on classification accuracy (y-axis) of changing training data size (x-axis) using stepwise MDA.

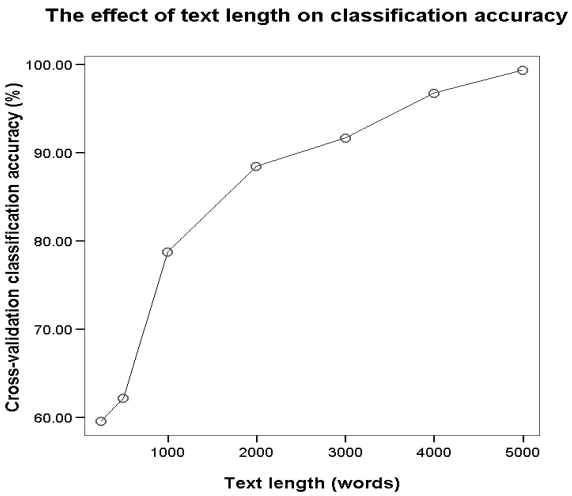

Figure 3: The effect of text length (x-axis) on classification accuracy (y-axis).

The plots show that classification accuracy is generally increasing in a nonlinear manner as the size of the training set increases until a plateau at of accuracy within the range 95-100%. The point of this plateau is at approximately 16 texts per author of classification model training data. These finding are consistent with those of Stamatos *et al.*[5] who concluded that in general, classification accuracy is improved by increasing the amount of training. Similarly, their findings show a plateau in accuracy at approximately 16 texts per author. Where these results differ from their findings is that they found lexically based approaches to be quite unstable in the way accuracy changes with changing training data size whereas this approach displays considerably more stability, in particular, by avoiding the significant reversals in the upward accuracy trend as training data size is increased. In addition, the technique developed here exhibits higher classification accuracy for all sizes of training data than reported by Stamatos *et al.*[5].

The second test (Figure 4b) shows that the classification accuracy increases at a decreasing rate as the text lengths are increased. The horizontal axis does not start at zero leading to an interesting, less obvious, result. Even with as few as 250 words per text, classification accuracy of close to 60% is achieved. By chance alone, with six authors in the corpus, an unbiased classification procedure should get one in six classifications right, that is, a classification accuracy of ≈16.7%. Hence classification accuracy of 60% is an improvement on chance of 43% with as few as 250 words per text. These results also suggest that to achieve a high classification accuracy of around 90% with six potential authors and having access to 26 texts per author, text length of approximately 2,000 to 3,000 words is required.

4.3. Application to the New Testament texts

4.3.1. Koine Greek New Testament

Most biblical scholars agree that the original text of the New Testament was written in Koine Greek, an ancient Greek dialect, with a minority considering an Aramaic language version to be the original. Consistent with this majority belief the Koine Greek sources of the New Testament[9] were used initially to eliminate the effects of translation on style characteristics. A review of current scholarly opinion on the authorship of the 27 books of the New Testament

identified 11 (in addition to *The Letter to the Hebrews*) as being of disputed or unknown authorship and eliminated from the corpus of training data. Two of the remaining books, *Philemon* and *Jude*, were also eliminated from the corpus due to their short length. Of the remaining 13 books, the ones of length greater than 6,000 words were split into shorter texts of approximately 3,000 words each in order to increase the size of the training set at the trade-off of text length. This was done in order to maximise overall classification accuracy due to an increase in accuracy from the increased number of texts in the training data exceeding the decrease in accuracy from the shortened average text length.

This corpus of texts was put through the style marker extraction process and then stepwise MDA. The plot below shows the scores of the 37 texts and *Hebrews* in the first two discriminant functions.

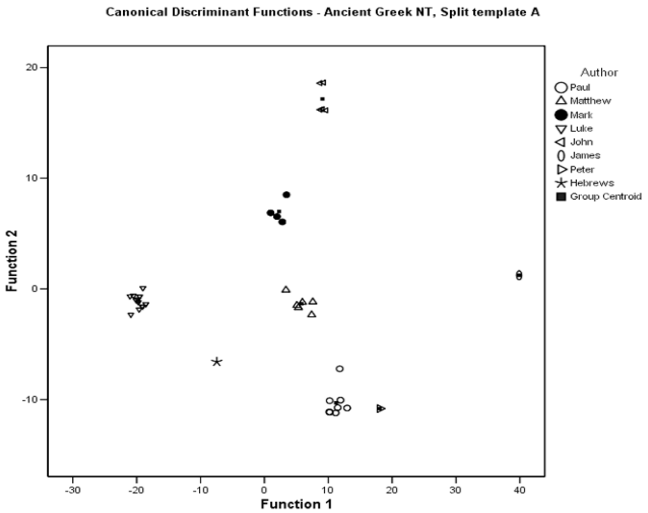

Figure 4: Scores of the texts in the Koine Greek New Testament corpus in the first two discriminant functions.

The plot only shows the first two of four discriminant functions—only part of the classification model. Further discrimination ability that separates groups is contained in the other two dimensions not seen in the plot. Also, such plots make texts of unknown authorship appear further from the centroids of potential authors than they would be if it were known that they belong to that particular group. This occurs because texts of known authorship have a bearing on the position of their group's centroid when constructing the classification model whereas texts of unknown authorship have no effect on the centroids.

In the plot the groups (authors) form clearly identifiable clusters. This is also reflected by the 100% LOO cross-validation classification accuracy and indicates a robust, accurate classification model. Using this model the Mahalanobis distances to centroids suggest *Hebrews* is most similar in authorial style to the texts attributed to Paul. However, the distance of *Hebrews* to Paul's centroid compared to the distance of Paul's texts to Paul's centroid suggests a significant amount of dissimilarity between Paul's texts and *Hebrews* motivating the search for other potential authors. Barnabas is one such potential author. The availability of an English translation of a text written by Barnabas as well as the ability to examine the effect of translation on style characteristics motivate the use of the King James Version of the New Testament for further analysis..

4.3.2. King James Version

At least one previous study has found that an author's style characteristics, particularly function word usage, are still present in translated texts[10]. It has been suggested by many scholars based on qualitative evidence that Barnabas (who is traditionally not attributed to the authorship of any of the books of the New Testament) wrote *Hebrews*. However, to the best of the authors' knowledge, this has never been quantitatively examined through the use of stylometric techniques. A text attributed to Barnabas, *The Epistle of Barnabas*, sourced as online electronic texts[11] was added to this corpus of texts. The same texts were eliminated from this corpus as were from the Koine Greek corpus. Again, the long texts were split to increase the size of the text corpus thus increasing overall classification accuracy.

Using this corpus of texts 70 single function word frequency stylometric vectors were extracted and a classification model constructed using stepwise MDA. The plot of discriminant function scores for this model is shown below.

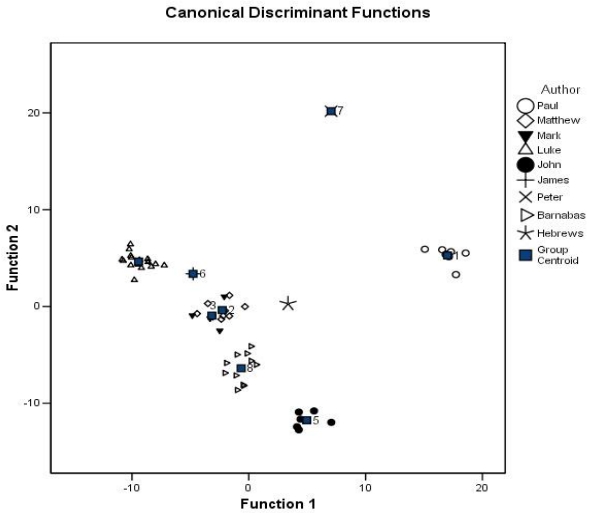

Figure 5: Scores of the texts in the King James New Testament text corpus in the first two discriminant functions.

The LOO cross-validation resulted in a classification accuracy of 100% indicating a robust and accurate classification model. Further this implies that authorial style characteristics have been preserved through translation to the extent that the stylometric markers of single function words are capable of discriminating well between authors. Using this, the most complete model, *Hebrews* is classified as being most stylometrically similar to the texts attributed to Barnabas. Interestingly this has been suggested by many scholars in past studies not employing stylometric evidence.

Mahalanobis distance is a powerful measure of the goodness-of-fit of a case characterized by a multivariate observation to a group's mean or centroid values. When comparing Mahalanobis distances of texts to centroids, the problem of a bias that appears in the plots was overcome by using the LOO cross-validation distances of texts to their centroids. The Mahalanobis distances were adjusted to account for the difference in distances from a text to its centroid caused by the splitting of texts that was undertaken to increase the size of the corpus. This can be though of as accounting for the fact that subsets of any single text will be more similar in style to one another than distinct texts written by the one author. The distribution of these distances was found to be approximately normal using one-sample Kolmogorov-Smirnov tests. Having only samples by which to estimate the unknown population mean and standard deviation of text distance to group centroid, it was necessary to use the Student's t-distribution when comparing distances of *Hebrews* to centroids with distances of texts of known authorship.

In making these comparisons it is important to note that it is not possible to conclude on stylometric analysis alone that a particular text was written by a particular author but it is possible to conclude that a particular text was not written by a particular author at a level of statistical significance. This is because there is a large number of authors that are possible in this multi-class problem, and stylometric analysis does not provide a unique signature similar to DNA or fingerprints. In a large sample of authors there will inevitably be some that overlap or share many stylistic features. This impacts the way the hypotheses being tested are written:

$$H_o: d_i \leq \mu_i \quad \text{vs.}$$
$$H_a: d_i > \mu_i$$

where H_o is the null hypothesis, H_a the alternate hypothesis, d_i is the distance of *Hebrews* to author *i*'s centroid and μ_i is the mean distance of author *i*'s texts to author *i*'s centroid. This one-tailed hypothesis test is equivalent to:

H_o: *Hebrews* MAY have been written by author *i* vs.
H_a: *Hebrews* was NOT written by author *i*.

In other words if the distance of *Hebrews* to a group centroid is less than or close to mean distance of the texts in that group to their centroid, the null hypothesis cannot be rejected and no conclusion can be made other than *Hebrews* may have been written by that author. If however, the distance of *Hebrews* to a group's centroid is extremely large relative to the group's mean distance having taken into consideration the standard deviation of the distances then the null hypothesis is rejected and the conclusion made that *Hebrews* was not written by that author at a level of statistical significance. This level of statistical significance depends on how extreme the distance of *Hebrews* to the centroid is. The further *Hebrews* is from the group's centroid relative to the texts in that group, the higher the confidence that *Hebrews* does not belong to that group.

The result of the hypothesis testing was that the null hypothesis, H_o, was rejected at a statistical significance level of 1% for the authors Paul, Matthew, Mark, Luke and John. The p-value, indicating the probability of observing a distance at least as extreme as the distance of *Hebrews* given that H_o is true, was as low as 8×10^{-20} in the case of John and the highest p-value was for Paul at 0.8%. The conclusion is made that there is very strong statistical evidence (greater than 99% confidence level) that *Hebrews* was not written by Paul, Matthew, Mark, Luke or John.

In the case of the author Barnabas, *Hebrews* was actually closer to the centroid of Barnabas than the mean of Barnabas's texts themselves. It is not possible for the distance of *Hebrews* to Barnabas's centroid to be systematically lower than Barnabas's mean distance (this can only happen by chance). Hence, without even having to perform the t-test on this distance, the null hypothesis, H_o, is unable to be rejected in the case of Barnabas. As a matter of interest, the p-value for this test is approximately 70% indicating that the observation is in fact not unlikely or extreme at all. The conclusion is made that *Hebrews* may have been written by Barnabas, who is the most likely of all authors examined to have written *Hebrews* (followed by Paul as distant runner-up).

5. CONCLUSIONS

Both of the first two authorship attribution techniques, multiple discriminant analysis of function word frequencies and the trigram Markov method, demonstrated the ability to correctly classify texts by author given a training set. Hence both techniques could be used to provide statistical evidence in authorship attribution problems. The third technique based on word recurrence intervals, although it may have some use if further developed, generated results too inconsistent for it to provide meaningful evidence in its current form.

The technique of multiple discriminant analysis of function word frequencies gave a best classification accuracy of 100% on the corpus of 156 fictional English texts. Classification accuracy very close to 100% was able to be consistently maintained even with changes to the training corpus, the number of stylometric markers used and the text length within limits. The trigram Markov method gave a best classification accuracy of 88.3% on the same English fictional text corpus as used in the first technique. This was achieved using 12 texts per author as training data. The accuracy is expected to increase with a larger sized set of training data, however, even a classification accuracy of 88.3% is good relative to accuracies reported by other studies.

The following conclusions can be made about statistical authorship attribution in general.

i) Classification accuracy increases as the number of stylometric markers (capable of characterising authorship) increases to a certain point after which model overfitting begins to reduce classification accuracy.
ii) Single function word frequencies are more capable of characterising authorial style than frequencies of sequences of function words.
iii) Classification accuracy generally increases as the size of the training data set is increased.
iv) Classification accuracy generally increases as the number of potential authors of a text of unknown authorship is decreased.
v) Classification accuracy generally increases as the average length of texts being used is increased.

The following conclusions can be made about who wrote *The Letter to the Hebrews*:

i) It is extremely unlikely that that *Hebrews* was written by Paul, Matthew, Mark, Luke or John. Statistically speaking, one can have greater than 99.1% confidence that *Hebrews* was not written by any of the five authors just mentioned.

ii) *Hebrews* is stylistically very similar to texts written by Barnabas. *Hebrews* may have been written by Barnabas.

REFERENCES

1. Love, H., *Attributing Authorship: An Introduction*, Cambridge University Press, 2002
2. Keselj, V., Peng, F., Cercone, N., & Thomas, C., "N-gram-based author profiles for authorship attribution," *Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING'03*, Dalhousie University, Halifax, Nova Scotia, Canada, August 2003.
3. Holmes, D. I., "Authorship attribution," *Computers and the Humanities,* **28**(2): 87-106, 1994.
4. Juola, P. & Baayen, H., "A controlled-corpus experiment in authorship identification by cross-entropy," *Literary and Linguistic Computing*, 2005 (In press).
5. Stamatatos, E., Fakotakis, N., & Kokkinakis, G., "Computer-based authorship attribution without lexical measures," *Computers and the Humanities*, **35**(2):193-214, 2001.
6. Karlgren, Jussi & Cutting, Douglass, Recognizing text Genres with Simple Metrics Using Discriminant Analysis, *Proc. of the 15th International Conference on Computational Linguistics (COLING '94)*, 1071-1075, 1994.
7. Zhao, Y. & Zobel, J., *Effective and Scalable Authorship Attribution Using Function Words*, RMIT University, Melbourne, Australia, 2005.
8. Project Gutenberg, <http://promo.net/pg/>, accessed 2005.
9. Nestle, E., Aland, K., Black, M., Martini, C., Metzger, B. & Wikgren, A. *Novum Testamentum Graece,* 26th edition, Deutsche Bibelgesellschaft, Stuttgart, 1979.
10. Farringdon, Jill, *Analysing for Authorship: A Guide to the Cusum Technique*, Cardiff: University of Wales Press, 1996.
11. *Early Christian Writings*, <http://www.earlychristianwritings.com/barnabas.html>, accessed 2005.
12. Khmelev, D. V., "Disputed authorship resolution through using relative empirical entropy for Markov chains of letters in human language text," *Journal of Quantitative Linguistics*, **7**(3): 201-207, 2000.
13. Oberlander, J. & Brew, C., "Stochastic text generation," *Philosophical Transactions of the Royal Society of London*, Series *A*, **358**(1769): 1373-1385, 2000.
14. Khmelev, D. V. & Tweedie, F. J., "Using Markov chains for identification of writers," *Literary and Linguistic Computing,* **16**(4): 299-307, 2001.
15. Ortuño, M., Carpena, P., Bernaola-Galván, P. Muñoz E., and Somoza1 A.M., "Keyword detection in natural languages and DNA," *Europhysics Letters,* **57**(5): 759-764, 2002.