A New Stylometric Technique for the Investigation of the True Authorship of the Early Works of Shakespeare

STEVEN JOHN WARDLE
SUPERVISOR PROFESSOR DAVID LOWE

MSc by Research in Pattern Analysis and Neural Networks

THE UNIVERSITY OF ASTON IN BIRMINGHAM

August 1997

THE UNIVERSITY OF ASTON IN BIRMINGHAM

A New Stylometric Technique for the Investigation of the True Authorship of the Early Works of Shakespeare

STEVEN JOHN WARDLE
SUPERVISOR PROFESSOR DAVID LOWE

MSc by Research in Pattern Analysis and Neural Networks, 1997

Thesis Summary

This thesis follows the development of a new stylometric technique which can be used to investigate works of literature of questionable authorship. A particular dispute is examined concerning the true authorship of three plays commonly credited to William Shakespeare. The works are the two historical plays Henry VI parts 2 and 3 and the Roman tragedy Titus Andronicus. The three works have been argued to have been greatly influenced by the playwright Christopher Marlowe. In a problem which is prone to high noise and limited data availability, intelligent feature selection is an essential part of the classification process. This should lead to the identification of the most promising inputs to use in discriminating between the two authors. The use of both linear classifier and non-linear Neural Network based classifier models are investigated and their results compared.

Keywords: stylometry, principal components analysis, neural networks, Marlowe, literary detection.

Acknowledgements

I would like to express my gratitude to the people who have contributed their support during the production of this thesis. I would like to thank Professor David Lowe for the advice and experience he has passed on throughout the past year. Secondly, I would like to thank both Robert Matthews and Tom Merriam whose ideas have proved invaluable both during the research and in the production of the final thesis. I would also like to thank Tom for supplying me with all of the play texts used throughout the research.

Contents

List of Figures

List of Tables

1. Introduction

A popular area of research in the field of literary studies is the question of authorship of disputed works. Described by Matthews & Merriam (1994b) as the quantitative analysis of literary style, stylometry takes a statistical approach to the problem of disputed authorship. This thesis concentrates on the development of a new stylometric technique. Whilst a particular authorship problem is investigated, the technique can be applied to any authorship dispute.

With stylometry, a comparison is made between disputed texts and those known to have been produced by candidate authors. The textual features selected for the comparison are called discriminators and should reflect the style of writing. By taking discriminator measurements for a large number of texts produced throughout an author's career it is possible to build a statistical model of that author's style. Measurements for a disputed text can then be examined to see which model they most closely match and the text can be credited to the corresponding author.

Stylometry dates back to the middle of the 19th century (Matthews & Merriam, 1994a) when Augustus De Morgan first suggested the use of mathematics in resolving authorship disputes. It was only in the early 20th century, with the development of new statistical techniques, that the field of stylometry was pushed forward significantly. Two independent scholars, G. Udny Yule and George Zipf, being considered responsible for the breakthrough. Up until the introduction of computers in the early 1960s, stylometric research entailed a great deal of physical and mental labour. Texts of great length had to be processed manually and calculations performed by hand.

The application of computers to the field of Stylometry has led to the examination of several key and controversial authorship disputes. Mosteller & Wallace (1964) investigate the authorship of the Federalist Papers, a set of political texts written in the 18th century to persuade the citizens of New York to ratify the

Constitution. Holmes (1992) examines the book of Mormon concluding that it is in fact the work of only one author, supporting the views of sceptics of the Mormon religion. However, perhaps unsurprisingly, the largest number of authorship disputes surround the works of arguably the greatest playwright to have ever lived, William Shakespeare. Numerous claimants have been suggested as the true authors of some of Shakespeare's works ranging from Francis Bacon to Queen Elizabeth I (Myers, 1990). Previous stylometric work relating to Shakespeare will be examined in Section 1.3.

The development of neural network techniques has led to a recent revival in stylometric investigations. Their main advantage is the ability to easily model non-linear relationships in the data. Tweedie *et al.* (1994) use such techniques, previously unavailable to Mosteller and Wallace, to revisit the problem of the authorship of the Federalist papers.

1.1 Major Issues

One of the largest areas of uncertainty in the field of stylometry is the question of purity of texts. The texts used throughout this thesis were obtained via the internet. They could have been copied from published editions dating from the 19th to the 20th century, brought together from several original versions by the respective editor. Although the author's characteristic style should still be preserved in edited versions, there is a great deal of external noise introduced into the data. No uniform convention for the English language existed at the time of the writing of the texts. The lack of standard spelling conventions resulted in certain letters occurring at a much higher frequency than they do in present day texts and vice versa. The manner in which these spellings are interpreted by an editor together with the editorial protocols used constitute the main sources of the noise. The high noise content is one of the reasons for investigating Neural Network techniques which are widely known to perform well with such problems.

The authorship dispute investigated in this thesis is considered particularly troublesome due to a relative shortage of author characteristic data. This problem is examined in greater detail in Section 1.2. Intelligent feature extraction therefore plays an important part in the classification process. Discriminators should be chosen so that

their use will allow the generation of as many samples as possible of the works of each author. However, it is also important that the discriminators contain enough information so as to reflect the characteristic style of the respective authors.

Finally, the stylometric technique selected for this thesis does make the problem one of very high dimensionality as identified in Section 1.4. This dimensionality must be reduced if a realistic model is to be constructed. This is a problem of feature selection. Only discriminators which best describe the differences between the authors should be used and the others discarded. Indeed, a significant part of this thesis is concerned with issues of how the large number of potential features can be reduced down to a small but characteristic set suitable for the specific problem under investigation.

1.2 Shakespeare and Marlowe

William Shakespeare and Christopher Marlowe were both born in the year 1564. Although it is believed that the two authors never met, a popular theory is that Marlowe had a strong influence on the young Shakespeare who paid tribute to him in the play *As You Like It* as the 'dead shepherd' (Dear, 1986). An alternately held view is that the works of Shakespeare are totally independent of those of Marlowe and that any similarity is coincidental.

The works that are considered to have the greatest Marlovian characteristics are the earliest works of Shakespeare. In particular the two historical plays, *Henry VI part 2* and *Henry VI part 3* and the Roman tragedy, *Titus Andronicus*. A strong literary relationship is believed to exist between the final two parts of *Henry VI* and two shorter anonymous works entitled *The Contention* and *The True Tragedy* (Matthews & Merriam, 1994b). A number of theories exist to explain this relationship and one actually credits the anonymous plays to Marlowe and suggests that Shakespeare adapted these to produce *Henry VI* parts 2 and 3.

Canons exist for both authors. These are sets of undisputed works attested by scholars to have been undoubtedly produced by an author. Matthews & Merriam (1994b) identify the contents of each canon. By analysing these plays, discriminators can be extracted which can be used to distinguish between the works of the two

authors. Whilst there are a large number of core canon plays for Shakespeare this is not the case for Marlowe. There are only three plays in the Marlowe core canon, *Tamburlaine I*, *Tamburlaine II* and *Edward II*. Doubts have been cast over the remaining plays credited to Marlowe.

Dido, full title *The Tragedie of Dido Queene of Carthage*, is believed to have been written whilst Marlowe was still at Cambridge University in collaboration with Thomas Nashe (Dear, 1986). Merriam (1995) argues that *The Jew of Malta* has been falsely credited to Marlowe and may have been written by Thomas Kydd. *Doctor Faustus* is thought to have been written towards the end of Marlowe's career and is commonly believed to contain a great deal of non Marlowe material added after the author's death. Several versions of the play actually exist and for this thesis the version first published in 1604 was used as this is now believed to contain fewer external additions although this is a questionable area itself (Ule, 1982). Ironically, the most accepted Marlowe play of the three is a fragmentary text, *Massacre at Paris*. Marlowe also produced some narrative poems including the erotic poem, *Hero and Leander*.

The uncertainty surrounding the true authorship of the plays outlined above introduces further noise into the data and obviously leads to a reduction in the amount of author characteristic data necessary for training a classifier.

1.3 Previous Work

Much research has been carried out into authorship questions regarding William Shakespeare. Matthews & Merriam (1994b) and Lowe & Matthews (1995) investigate the authorship of the play *The Two Noble Kinsmen*. The frequencies at which certain words occur were chosen as discriminators. The results of both investigations support the claim that the play was a collaboration between Shakespeare and his contemporary John Fletcher.

In the investigations referenced above, function words were chosen as discriminators. The frequency of function words such as 'the', 'as' and 'may' would be very difficult to imitate deliberately in the case of forgery unlike the frequency of more

exotic words. However, Udny Yule (1968) successfully used the frequencies of such exotic words as discriminators in a number of investigations.

Ledger & Merriam (1994) also examine the authorship of *The Two Noble Kinsmen*. They demonstrate that letter frequencies can be used as an alternative to word frequencies. As with the use of function words, it would be very difficult to consciously influence letter frequency counts whilst attempting to imitate the work of another author. Also with an average of 5 - 6 letters per word, any sample of text will produce more information when using letter frequencies than when using word frequencies. This should lead to a lower volume of text being required to train a classifier model or, with higher counts available for a given text length, better statistics for the chosen discriminators. Word and letter frequencies are not the only measurements used in stylometry. Holmes (1994) examines a number of other methods that have been used.

Regarding discrimination between Shakespeare and Marlowe, Merriam (1996) uses the frequencies of both function words and exotic words to show that seven historical Shakespeare plays, including the second and third parts of *Henry VI*, as well as the tragedy *Titus Andronicus* have some Marlovian characteristics. Matthews & Merriam (1994b) use function words to support the argument that *The Contention* and *The True Tragedy* are both works by Marlowe adapted by Shakespeare to produce the second and third parts of *Henry VI*.

1.4 Approach to be Followed

For this thesis a new stylometric unit will be introduced called a *digram*. A digram describes a pair of letters occurring consecutively in a word. For example, the word 'author' is made up of the five digrams 'au', 'ut', 'th', 'ho' and 'or'. Punctuation and space characters are not used in the formation of digrams. Also, word concatenation, where the last letter of one word and the first letter of the following word could form a digram, is not employed. The frequencies of selected digrams can be used as discriminators in a similar way to those of single letters (Ledger & Merriam, 1994). As with single letters, it would be very difficult for an author to consciously affect digram counts whilst trying to imitate another author.

Another similarity to single letter frequencies is that a lower text volume is necessary to build a classifier model than when using word frequencies. This has already been highlighted in Section 1.2 as a characteristic of the particular authorship dispute being investigated. Ideally a technique is required which can be used with very limited volumes of text. One possible use for such a technique would be in resolving claims that police statements have been falsified in criminal cases (Bailey, 1979). In such cases the disputed texts may be no longer than half a page in length. Whilst a technique which could be used successfully in such circumstances may never be developed, the need for a technique based on low volumes of text is indicated.

A total of 676 digrams are available in the English language ranging from 'AA' to 'ZZ'. It would not be feasible to use the frequencies of these to create a 676 variable input space. Some method of selecting a lesser amount of discriminators must be found. The new discriminators should contain as much information as possible so that an adequate model can be constructed using them. Principal Components Analysis (PCA) is a method of reducing dimensionality whilst still explaining a large amount of the variance in a data set (Section 2.1). PCA will be used in two ways. Firstly, to identify suitable digrams to act as discriminators in a new feature space. Secondly, the results of PCA will be combined with digram frequencies to construct totally new discriminator variables.

Neural Network techniques (Section 2.3) can be used to model non-linear relationships in the new feature space. However, if the relationships are simply of a linear nature then a linear model (Section 2.2) would be adequate. Therefore, linear models will be used to produce benchmark results against which the performance of Neural Network based classifiers can be compared.

In addition to linear models, neural network based models will be applied to the authorship problem concerning Shakespeare and Marlowe. The results produced by all models will indicate the suitability of digrams as discriminators in authorship classification problems.

2. Theory

This section contains an outline of the various statistical techniques used as part of the stylometric analysis.

2.1 Principal Components Analysis

Alt (1990) provides a good non-mathematical description of Principal Components Analysis (PCA) and its uses. PCA attempts to describe the variance in a data set using a smaller number of variables called principal components. The principal components are actually the eigenvectors derived from the covariance matrix generated from the data set. Each eigenvector has a corresponding eigenvalue which indicates the importance of the associated principal component in the description of the overall variance. The following formula can be used to determine the percentage of the total variance in a data set that is described by the first k ordered principal components

$$\text{Percentage Variance Described} = \sum_{i=1}^{k} \lambda_i \Big/ \sum_{i=1}^{N} \lambda_i \times 100\ \% \qquad \textbf{(2.1)}$$

where λ_i is the *ith* eigenvalue and N is the total number of eigenvalues.

Singular Value Decomposition (SVD) is a simple and effective method of performing PCA and was the technique used in this thesis. If the data set A is an ($n \times p$) matrix of rank r then it can be written as follows

$$A = USV' \qquad \textbf{(2.2)}$$

where U ($n \times p$) and V ($p \times p$) are column orthonormal matrices and S is a diagonal matrix. The columns of U are actually the eigenvectors of AA' and the columns of V are the eigenvectors of A'A. It can be shown that the values on the diagonal of S, called singular values, are the square roots of the eigenvalues corresponding to the derived eigenvectors. A matrix of rank r produces only r significant singular values. As eigenvalues simply provide a measure of the contribution of each principal component, singular values can be used interchangeably with eigenvalues throughout this thesis.

2.2 Linear Classifiers

Three different linear classifiers were utilised in the work. One particular model consistently produced superior results, the 'Linear Network' classifier, and the outputs produced by this model are the only results quoted for linear classifiers. All three models will be described briefly in this section.

A training data set consists of n patterns of class A and n patterns of class B. Each pattern is represented by a vector of d variables and is said to be d-dimensional. With a 'Nearest Class Mean' classifier model, two vectors are constructed. One by calculating the mean of all vectors representing patterns from class A and the other by calculating the mean of all vectors representing patterns from class B. The Euclidean distance between each mean vector and a vector representing an unclassed data pattern can be measured and the new pattern allocated to the class corresponding to the shortest distance. The Euclidean distance, d, between two vectors **x** and **y** each containing n elements is calculated using the following formula

$$d = \sqrt{\left((x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2 \right)} \qquad \textbf{(2.3)}$$

With a 'Nearest Neighbour' classifier model, the Euclidean distances between the vector representing an unclassed data pattern and the individual vectors representing each of the patterns in the training data set are measured. The pattern corresponding to the shortest distance is deemed to be the nearest neighbour of the new pattern and the new pattern is allocated to the same class.

A 'Linear Network' uses a set of weights, W, such that

$$AW = T \qquad (2.4)$$

where A is the training data set and T is a set of binary target value pairs, [0,1] for class A and [1,0] for class B. This equation can be rearranged to calculate the values for the network weights

$$W = A^{+}T \qquad (2.5)$$

where A^{+} is the pseudo-inverse of the matrix A. By multiplying a vector representing an unclassified pattern by the generated weights matrix, a pair of outputs will be produced. The new pattern can then be allocated to the class corresponding to the most similar target pair.

2.3 Non-Linear Classifiers

The models described so far are only capable of fully solving problems of a linear nature. Neural Networks are classifiers capable of modelling non-linear relationships in a data set. Both Haykin (1994) and Bishop (1995) provide a good introduction to neural network techniques and their application. Two types of neural network were used in the production of this thesis and both are outlined below.

2.3.1 Multi-Layer Perceptron networks

Figure 2.1 contains an illustration of a Multi Layer Perceptron (MLP) network. It consists of a number of input nodes, hidden nodes and output nodes. Each node has several inputs from nodes in the previous layer as well as an additional input called the bias which represents the overall contribution of the node. These inputs are multiplied by weights associated with the relevant connections and the sum of the resulting values is calculated. This sum is calculated for an arbitrary node, q, using the following formula

$$S_q = \sum_{p=0}^{N} w_{pq} x_p \qquad \textbf{(2.6)}$$

where x_p is the p*th* input to the node, w_{pq} is the weight associated with the connection between input p and node q and the sum runs over all N input nodes. The input x_0 is permanently held at 1 and the bias is implemented by setting w_{0q} accordingly.

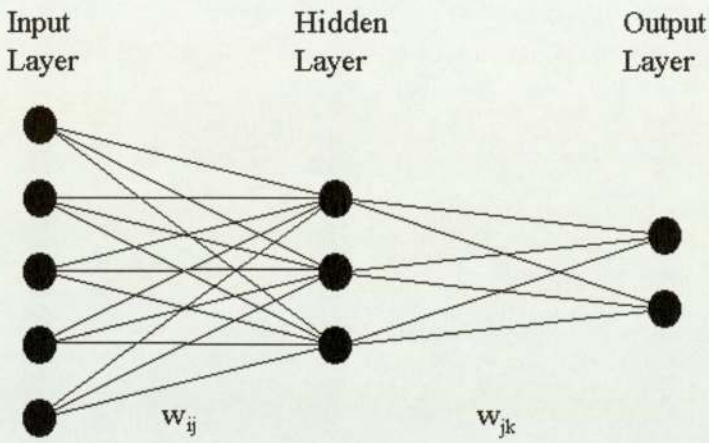

Figure 2.1 Structure of a Multilayer Perceptron Network

This sum is then passed through a transfer function and the node output is defined as the value of the transfer function. For a hidden layer node, q, a sigmoidal transfer function is used as follows

$$O_q = 1 / 1 + \exp(-S_q) \qquad \textbf{(2.7)}$$

Although the output of a node, q, in the output layer of an MLP network can also be defined using a non-linear transfer function such as the one given above, for this work a linear transfer function such as the one below was used.

$$\begin{aligned} O_q &= 0 && \text{when } S_q < 0 \\ &= 1 && \text{when } S_q > 1 \\ &= S_q && \text{otherwise} \end{aligned}$$

Whereas the training of a Linear Network is a single pass process, the weights connecting the hidden layer to the input and output layers in an MLP network are modified a number of times as the network is trained.

During training, a vector representing a pattern from a training data set is presented to the network together with a target vector. After the propagation of the input vector through the network, the network weights are modified so as to try to minimise the error between the actual network output and the target output. One algorithm for doing this is the Scaled Conjugate Gradient (SCG) algorithm which is described in detail in Bishop (1995). Apart from being faster than some traditional methods such as Gradient Descent, this method is particularly effective as it tends to avoid termination at local minima on the error surface. The weights of all MLP networks used in the production of this thesis were initialised to random values between -1 and 1 and then the networks were trained using the SCG algorithm.

One important decision is when to terminate the training process. If training is terminated too early, the network will not have had the opportunity to model the data as closely as it might. Conversely if termination is left too late, the network may be modelling the data almost exactly including any noise that will undoubtedly be present, a condition known as 'over-training'. In both situations, the network will not be able to generalise well from patterns in the training data to classify previously unseen patterns.

There are several methods available to determine when the training process should be terminated. For this thesis a relatively simple method was selected. The network is trained on each pattern in the training data set individually. Multiple iterations through the training set are used. After each complete iteration, the overall network error on a separate validation set is calculated. If this is greater than the validation error calculated at the end of the previous iteration then the overall generalisation ability of the network has deteriorated and training is terminated.

2.3.2 Radial Basis Function networks

The Radial Basis Function (RBF) network has a similar structure to an MLP network, consisting of an input, hidden and output layer. As with the MLP network outlined previously, the nodes in the output layer utilise linear transfer functions.

However, each hidden node represents a 'centre' in the feature space. By placing a basis function around each centre, a new pattern vector can be classified according to its distance from each centre.

One major difference between MLP and RBF networks concerns how the first layer weights are interpreted. With an MLP network, the products of the inputs and first layer weights are calculated, summed and passed through a transfer function. However, in an RBF network, the differences between the inputs and weights are used. The output of a hidden node, q, is defined as the value of some non-linear transfer function. Throughout this work the Gaussian transfer function given below was used.

$$\varphi_q(\mathbf{x}) = \exp [(\| \mathbf{x} - \boldsymbol{\mu}_q \|) \Sigma^{-1} (\| \mathbf{x} - \boldsymbol{\mu}_q \|)] \qquad \textbf{(2.8)}$$

where $\mathbf{x}$ is the input vector and $\boldsymbol{\mu}_q$ is a vector determining the centre of the basis function associated with hidden node q. This value can be used to calculate the output of a node, r, in the hidden layer using the following formula

$$O_r = \sum_{q=0}^{N} w_{qr}\varphi_q(\mathbf{x}) \qquad \textbf{(2.9)}$$

where w_{qr} is the weight associated with the connection between hidden node q and output node r. Again the bias is implemented by holding φ_0 at 1 and setting the value of w_{0r} appropriately.

There are a number of suitable methods for training an RBF network. The networks used for this thesis were trained in a two-stage process. In the first stage, the basis function centres are placed at vectors corresponding to randomly selected patterns in the training set. These centres are then kept fixed while the second layer weights are determined in the second stage by following the process detailed in Bishop (1995).

Although similar in function, RBF networks have a number of advantages over MLP networks especially when used with problems involving low volumes of sample data. A finite size data set has a finite number of degrees of freedom. The optimisation

of network parameters (weights and biases) use some of these degrees of freedom. Clearly a large MLP network with too many parameters would require more degrees of freedom than are actually available in a small data set preventing successful optimisation.

With an RBF network, prior knowledge of the data set is used to initially place basis functions in areas of high data density resulting in the distribution of the data already being represented. The only optimisation necessary then involves the final layer weights. Therefore, fewer parameters need to be optimised than for an equivalent size MLP network, requiring a smaller amount of training data.

3. Experiments

3.1 Pre-processing of Data

Before any form of analysis can take place, data samples must be generated for each author. The texts are partitioned according to a selected window size, for example a window of 1000 digrams. For each partition the number of every possible digram is counted and converted to a frequency. By performing this process for a number of partitions, using various Marlowe and Shakespeare plays taken from various stages in the authors' careers, a data set is built up containing an equal number of samples for both authors. This training set can then be used to train a classifier model. Tweedie *et al.* (1996) suggest using the following formula to determine the number of samples required to train a neural network based classifier.

$$N_T = 10(N_I + N_O) \qquad \textbf{(3.1)}$$

where N_T is the minimum number of training patterns, N_I is the number of inputs and N_O is the number of outputs. This formula will be used as a guideline to select the necessary number of partitions when using both linear and neural network classifier models.

The core-canon plays from which a training data set is generated are listed in Table 3.1. One problem with the plays is that speech headings and stage directions are repeated throughout the text, possibly affecting the digram counts. These therefore need to be removed during the pre-processing stage. The figures in parentheses in Table 3.1 indicate the total number of remaining digrams in each of the plays after their removal.

Marlowe Training Data	Shakespeare Training Data
Tamburlaine I (58165)	Much Ado About Nothing (60675)
Tamburlaine II (58583)	Julius Caesar (57670)
Edward II (63808)	Romeo and Juliet (71175)
	Merchant of Venice (62099)
	Antony and Cleopatra (72645)
	Twelfth Night (56226)
	The Winter's Tale (74109)
	Henry IV part I (72055)

Table 3.1 List of core canon Marlowe and Shakespeare plays used to generate training data

In addition to the data used for classifier training, a separate data set is necessary containing previously unseen samples to test the generalisation performance of a classifier. The data samples for this validation set are generated from the core canon plays listed in Table 3.2. The Marlowe texts in Table 3.2 are identical to those in Table 3.1. This is due to the limited size of the Marlowe canon. It is therefore essential that the partitions used to generate the Marlowe samples in the validation set are different to those used for the training set. In reality, overlapping windows were used to enable the generation of a sufficient number of Marlowe samples to be used in both the training and validation data sets.

Marlowe Validation Data	Shakespeare Validation Data
Tamburlaine I (58165)	Comedy of Errors (41706)
Tamburlaine II (58583)	A Midsummer Night's Dream (48886)
Edward II (63808)	All's Well That Ends Well (67300)

Table 3.2 List of core canon Marlowe and Shakespeare plays used to generate validation data

During the final stage of pre-processing, the frequencies of all digrams are normalised to be of zero mean and unit variance. This ensures that the frequencies of all digrams contribute equally during the training process.

3.2 Most Common Digram Frequency Approach

As mentioned previously, the English language permits a total of 676 possible digrams. This is clearly too large a number to analyse conveniently. By eliminating all

digrams which do not occur throughout the play texts used to generate the training data, this number can be reduced to 473. Furthermore, by eliminating any digram which occurs less than once in every one hundred digrams throughout the training data plays, 25 digrams remain. These are listed in Table 3.3. It can be argued that digrams which occur at a frequency of less than 1% are more likely to be sensitive to a play's context. Noise would also have a greater effect on the frequencies of these rarer digrams.

Digram				
AN	ER	IN	ND	RE
AR	ES	IS	NO	ST
AT	HA	IT	ON	TH
EA	HE	LL	OR	TO
EN	HI	ME	OU	VE

Table 3.3 List of the digrams which occur at a greater frequency than 1% throughout the plays from which the training data is generated

Principal Components Analysis (PCA) can be used to investigate the variance in a data set containing the normalised frequencies of each of these 25 digrams for equal numbers of Marlowe and Shakespeare samples. Figure 3.1 contains a plot of the projection of the normalised frequencies for each sample onto the first two principal components using a 3000 digram sample size. The data set used contained 116 Marlowe and 116 Shakespeare samples generated from the plays listed in Table 3.1.

Appendix A contains similar plots constructed from data sets generated using different sample sizes. It was decided to continue with a sample size of 3000 digrams as Figure 3.1 shows a good initial discrimination using just the first two principal components. The use of a sample size of 3000 digrams also enables the generation of enough samples to satisfy formula 3.1.

Alt (1990) explains how to interpret the results of a Principal Components Analysis. PCA produces one eigenvector and one corresponding eigenvalue (or singular value) for each Principal Component. The singular values indicate how much of the overall variance in the data is being described by the principal component.

A common statistical technique can be employed to identify which digram frequencies contribute greatest to the underlying variance in the data. The singular value spectra produced after performing PCA on the training data samples is given in

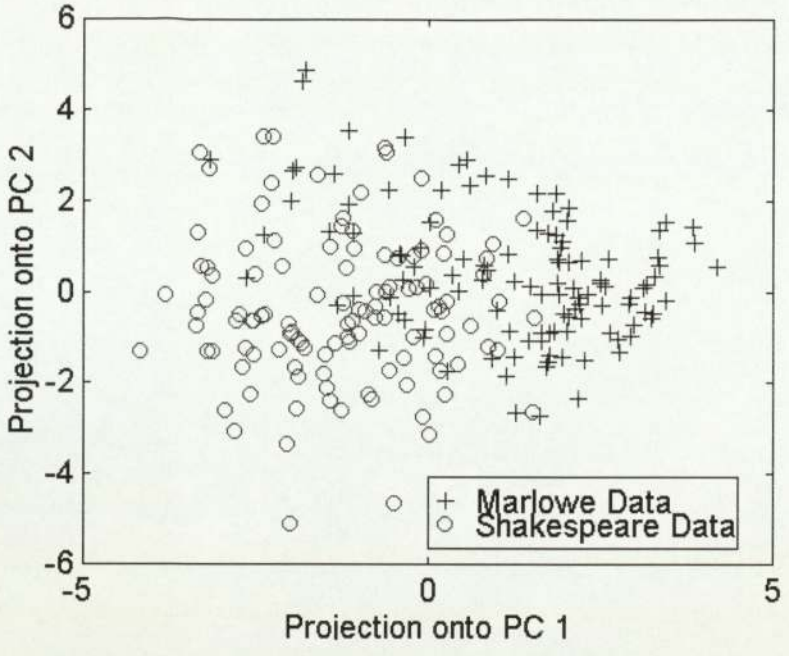

Figure 3.1 Projection of the 25 digram data set onto the first two principal components using a 3000 digram sample size

Figure 3.2. A closer examination reveals that the singular values appear to decline smoothly and exponentially for the first five principal components. After this, kinks appear in the spectrum. One interpretation is that the first five principal components are mostly describing the variance due to the differences between the two authors. The remaining principal components, that is principal components 6 to 25, are assumed to be mainly describing the noise in the data. By substituting the squares of the singular values for the eigenvalues in formula 2.1 it can be seen that the first five principal components are only responsible for approximately 44% of the total variance. With the remaining principal components describing the noise in the data, this figure supports the statement that the problem is one of high noise content.

Each eigenvector consists of one 'loading' for each variable, or digram, indicating how much that variable contributes to the direction of the corresponding principal component. The eigenvectors corresponding to the first five principal components are given in Appendix B. Digrams with high overall loadings over these eigenvectors can be assumed to be the greatest contributors to the construction of their corresponding principal components. As all other principal components are believed to be simply describing the noise in the data, these digrams must also contribute greatest to the underlying variance in the data set. This suggests that these

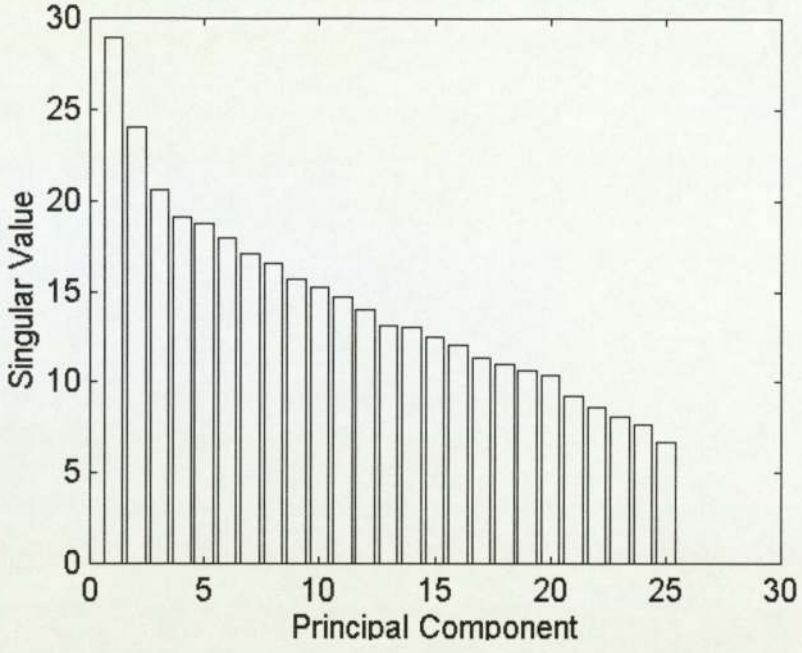

Figure 3.2 Singular Values obtained when PCA is performed on the 25 digram data set

Digram
AN
HA
HE
IS
ND
NO
OU
TH

Table 3.4 List of the 8 digrams which contribute most to the first five principal components of the 25 digram data set

would be the most suitable digrams to use as discriminators when distinguishing between the two authors. The eight digrams identified by this process are listed in Table 3.4.

A visual inspection of the normalised frequencies of each of the 25 digrams shows that nine digrams provide a strong discrimination between the Marlowe and Shakespeare data samples. This is illustrated in Figure 3.3. Each point represents a single data sample from the training set. Its position along the y-axis corresponds to the normalised frequency at which a particular digram, the digram 'ND' appears in that sample. The location along the x-axis simply represents the position of the sample in

the total 116 Marlowe or 116 Shakespeare samples in the training data set. It can be seen that in general, the normalised frequencies of the digram throughout the Marlowe samples are positive. Conversely, the frequencies of the digram throughout the Shakespeare samples appear to be mostly negative. The nine digrams identified by a visual analysis in this manner are listed in Table 3.5. Five digrams appear both in this new list and in Table 3.4. These digrams are listed in Table 3.6.

Digram
HA
IS
IT
ME
ND
NO
ON
OR
OU

Table 3.5 List of the 9 digrams which provide best discrimination by a visual inspection of digram frequency distributions

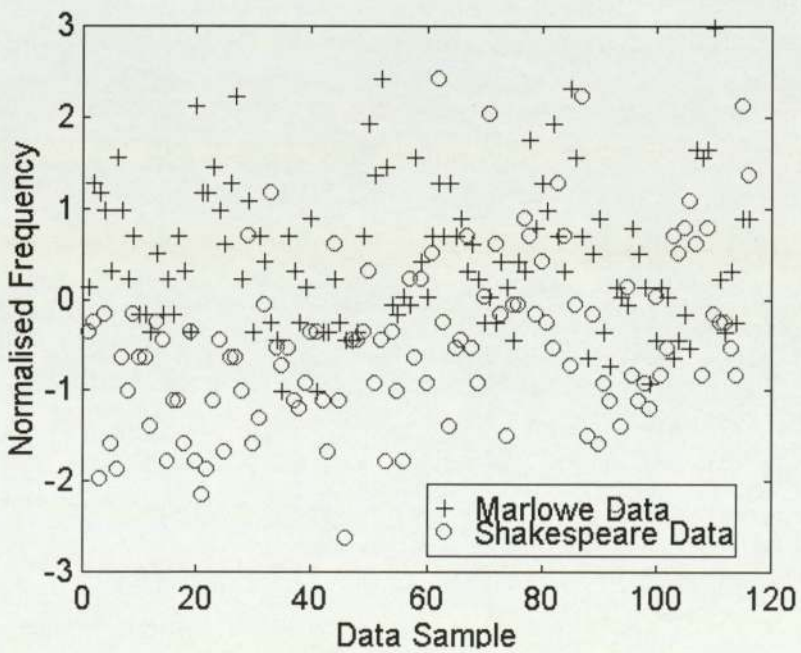

Figure 3.3 Normalised Frequencies of the digram 'ND' over 116 Marlowe and 116 Shakespeare data samples

Digram
HA
IS
ND
NO
OU

Table 3.6 List of the 5 digrams which provide best discrimination agreed by PCA and a visual analysis

Table 3.7 contains a combined list of the digrams identified by a visual analysis (Table 3.5) together with those derived by performing PCA (Table 3.4). The normalised frequencies of the digrams identified in Table 3.6 and Table 3.7 can be used as inputs to 5-input and 12-input classifier models respectively.

Digram
AN
HA
HE
IS
IT
ME
ND
NO
ON
OR
OU
TH

Table 3.7 List of the 12 digrams which provide best discrimination either by PCA or by a visual analysis

The normalised frequencies of the five digrams listed in Table 3.6 were presented to a linear network classifier model, as described in Section 2.2. The training set consisted of 116 Marlowe and 116 Shakespeare samples from the plays listed in Table 3.1. The results of testing the network on a validation set consisting of 30 Marlowe and 30 Shakespeare samples from the plays listed in Table 3.2 are displayed in Table 3.8 in the form of a confusion matrix. This compares the number of samples classified correctly to the number misclassified by the network.

	Predicted as Shakespeare	Predicted as Marlowe
Actual Shakespeare	20	10
Actual Marlowe	5	25

Table 3.8 Confusion Matrix obtained using Linear Network with 5 digram inputs on validation set

The results in Table 3.8 can be interpreted in an alternative way by considering the posterior probability. That is, the probability that an author is actually responsible for a sample given that the network has credited it to one of the authors. For example, the probability that Shakespeare actually produced a sample that the network has classified as Marlowe is equal to the number of Shakespeare samples misclassified as Marlowe divided by the total number of samples classified as Marlowe.

$$\begin{aligned} P(\text{Shakespeare} \mid \text{Marlowe Predicted}) &= 10 / (10 + 25) \\ &= 10 / 35 \\ &= 0.29 \end{aligned}$$

Table 3.9 lists the posterior probabilities derived from Table 3.8 using the above method. This table can be used to easily visualise the reliability of the classifications of the network. The calculation of posterior probabilities is commonly used in medical analysis to produce 'false positive' and 'false negative' percentages (Campbell & Machin, 1993). The classification of any sample as Shakespeare by a network might be viewed as a positive event. If so, the false positive percentage is equal to the probability that Marlowe produced a sample which the network has credited to Shakespeare. Conversely, the false negative percentage is the probability that Shakespeare actually produced a sample credited to Marlowe by the network.

P(Shakespeare | Shakespeare Predicted)	**0.8**
P(Shakespeare | Marlowe Predicted)	**0.29**
P(Marlowe | Marlowe Predicted)	**0.71**
P(Marlowe | Shakespeare Predicted)	**0.2**

Table 3.9 Posterior probabilities derived from the results of applying a Linear Network with 5 digram inputs to validation set

Table 3.10 displays the confusion matrix obtained by training a similar linear classifier using the frequencies of the twelve digrams listed in Table 3.7. The

corresponding posterior probabilities are given in Table 3.11. It can be clearly seen that a better performance is produced using twelve inputs. The results in Table 3.8 suggest that the frequencies of the five digrams used do not contain enough information to act alone as discriminators with 25% of the previously unseen patterns being misclassified.

	Predicted as Shakespeare	**Predicted as Marlowe**
Actual Shakespeare	26	4
Actual Marlowe	1	29

Table 3.10 Confusion Matrix obtained using Linear Network with 12 digram inputs on validation set

P(Shakespeare | Shakespeare Predicted)	**0.96**
P(Shakespeare | Marlowe Predicted)	**0.12**
P(Marlowe | Marlowe Predicted)	**0.88**
P(Marlowe | Shakespeare Predicted)	**0.04**

Table 3.11 Posterior probabilities derived from the results of applying a Linear Network with 12 digram inputs to validation set

The confusion matrices produced using the linear networks suggest that the data is not linearly separable into two distinct classes. A neural network model can be constructed to investigate whether better discrimination is possible using a non-linear technique. The structure of the most suitable network is determined by training MLP networks of increasing complexities to completion and plotting the final sum-of-squares error over the training and validation sets. The sum-of-squares error describes the difference between the target output and the actual output obtained from the network. It is calculated using the following formula

$$E = \sum_{n=1}^{N} \| \mathbf{y}_n - \mathbf{t}_n \|^2 / N \qquad \textbf{(3.2)}$$

where $\mathbf{y}_n$ is the network output vector, $\mathbf{t}_n$ is the target output vector and the sum runs over all N patterns in the training set.

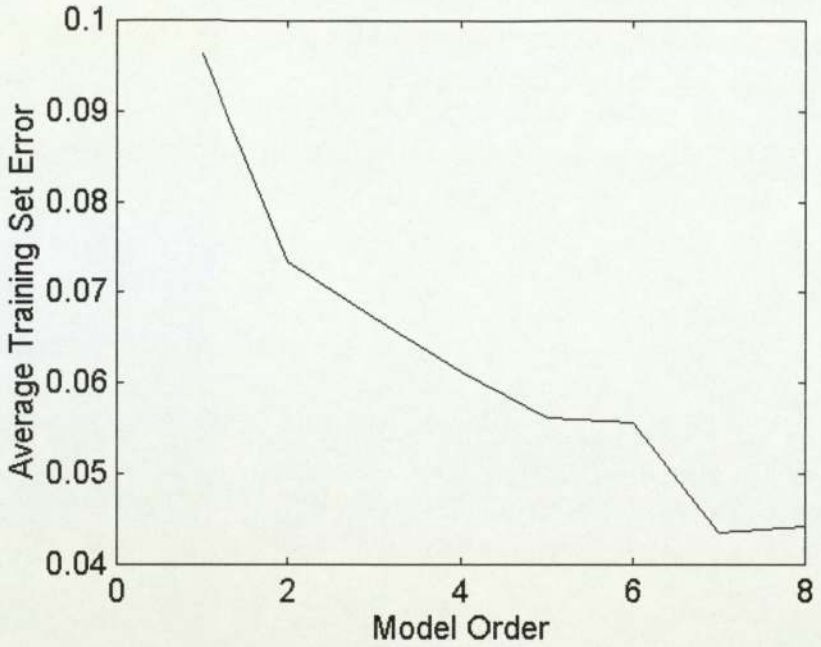

Figure 3.4 Training Set Error obtained after training MLP networks with different numbers of hidden units

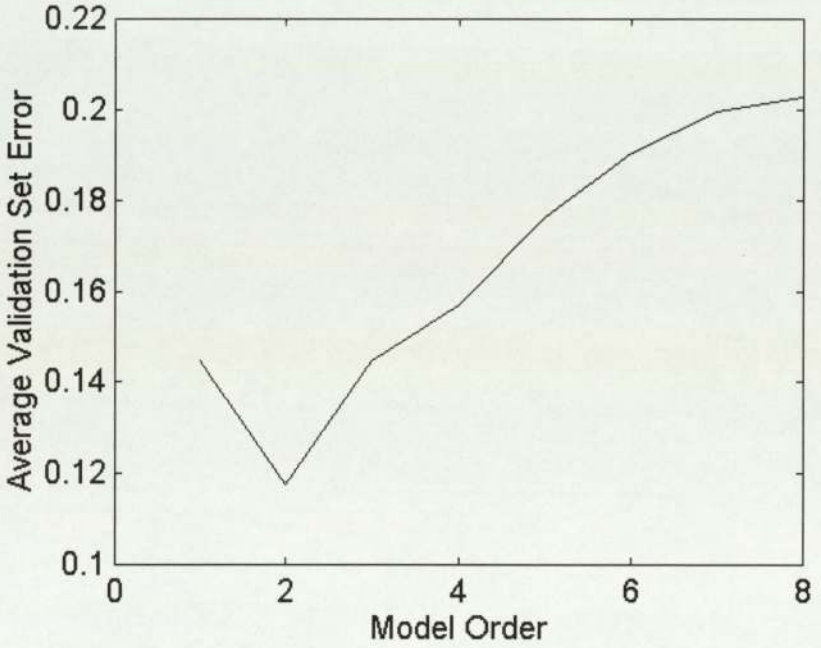

Figure 3.5 Validation Set Error obtained after training MLP networks with different numbers of hidden units

Figure 3.4 and Figure 3.5 illustrate the average error on the training and validation sets after training five different MLP networks for each number of hidden units. Figure 3.5 shows that generalisation performance deteriorates

when more than two hidden units are used. This indicates that an MLP model with two hidden units should produce the best results. The requirement for such a small number of hidden units suggests that an MLP network will produce very little improvement over a linear network.

Table 3.12 displays the confusion matrix obtained by applying an MLP network (as described in Section 2.3.1) with twelve inputs and two hidden units, trained on the samples in the training set, to the samples in the validation set. Table 3.13 gives the posterior probabilities calculated from the results in the confusion matrix. The results in both tables demonstrate that, as expected, the use of an MLP network offers very little advantage over a linear network with only one additional data sample being classified correctly.

	Predicted as Shakespeare	**Predicted as Marlowe**
Actual Shakespeare	26	4
Actual Marlowe	0	30

Table 3.12 Confusion Matrix obtained using MLP with 12 digram inputs and 2 hidden units on validation set

P(Shakespeare | Shakespeare Predicted)	**1.0**
P(Shakespeare | Marlowe Predicted)	**0.12**
P(Marlowe | Marlowe Predicted)	**0.88**
P(Marlowe | Shakespeare Predicted)	**0.0**

Table 3.13 Posterior probabilities derived from the results of applying an MLP Network with 12 digram inputs and 2 hidden units to validation set

3.3 Single Letter Frequency Approach

An alternative to using digrams is to use the frequencies of a subset of the 26 single letters as inputs to a classifier model. Figure 3.6 shows the projection of a data set containing the frequency of all 26 letters onto the first two principal components identified by PCA. The data set consists of 92 Marlowe and 92 Shakespeare samples generated from the plays in Table 3.1 using a sample size of 2500 letters.

PCA produces the singular value spectra illustrated in Figure 3.7. The values appear to decline smoothly and exponentially for the first three principal components

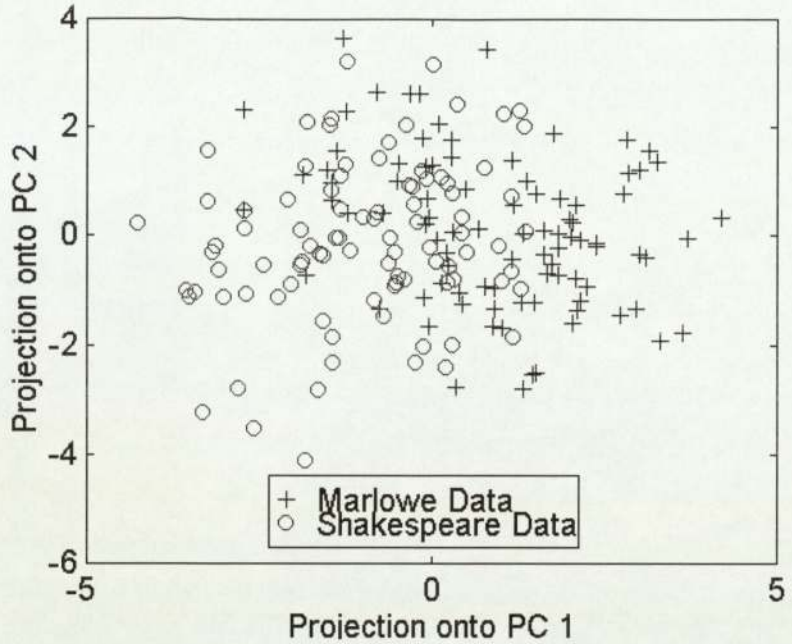

Figure 3.6 Projection of the 26 single letter frequency data set onto the first two principal components using a 2500 letters sample size

only. This would indicate that the remaining principal components are mainly describing the noise on the data. The use of single letter frequencies appears to be more susceptible to external noise than digram frequencies with only 26% of the overall variance being explained by the first three principal components.

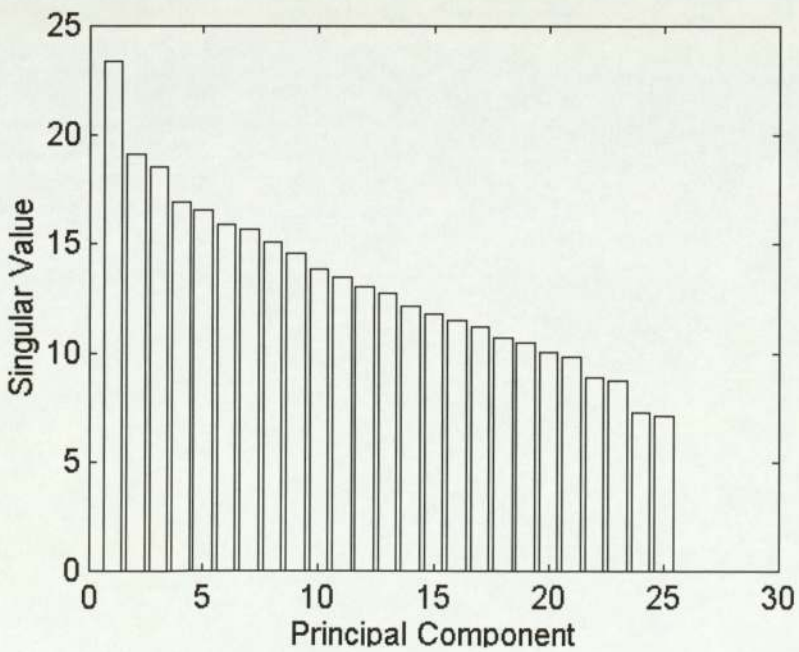

Figure 3.7 Singular Values obtained when PCA is performed on the 26 single letter data set

The eigenvectors corresponding to the first three principal components are given in Appendix B. A closer examination of this table identifies the frequencies of the letters listed in Table 3.14 as the variables which contribute most to the first three principal components. This would indicate that the frequencies of these letters should be used as discriminators. A linear network was trained on the 184 sample training set using the frequencies of these ten variables as inputs. The results of applying the network to a validation set generated from the plays listed in Table 3.2 are given in the form of a confusion matrix in Table 3.15 and as a list of posterior probabilities in Table 3.16.

Letter
E
F
H
K
O
R
S
U
W
Y

Table 3.14 List of Single Letters which provide best discrimination by PCA

	Predicted as Shakespeare	Predicted as Marlowe
Actual Shakespeare	15	15
Actual Marlowe	10	20

Table 3.15 Confusion Matrix obtained using a Linear Network with 10 single letter inputs on validation set

P(Shakespeare | Shakespeare Predicted)	**0.6**
P(Shakespeare | Marlowe Predicted)	**0.43**
P(Marlowe | Marlowe Predicted)	**0.57**
P(Marlowe | Shakespeare Predicted)	**0.4**

Table 3.16 Posterior probabilities derived from the results of applying a Linear Network with 10 single letter inputs to validation set

By comparing the results in both tables to previous results it can be seen that digram frequencies appear to contain more information than the frequencies of single

letters. This is particularly visible in Table 3.16 which shows that very little confidence can be placed in any network prediction when using single letter frequencies. For this reason, single letter frequency based classifiers will not be investigated any further.

3.4 Principal Component Projection Approach

The relatively poor performance of classifiers using the frequencies of five digrams as inputs is probably due to too much information being discarded in reducing the number of variables. A method of maintaining a large amount of information whilst limiting the number of input variables needs to be employed. This can be done by using the values obtained by projecting a data set onto the first few principal components as inputs to a classifier. To ensure that as much information as possible is used, PCA is carried out using all possible digrams and not just the most common ones. Figure 3.8 contains a plot of the projection of a data set onto the first two principal components identified by PCA. The training data set contained the normalised frequencies of all 473 occurring digrams for each of 117 Marlowe and 117 Shakespeare samples generated from the training data plays. Once again a sample size of 3000 digrams was used.

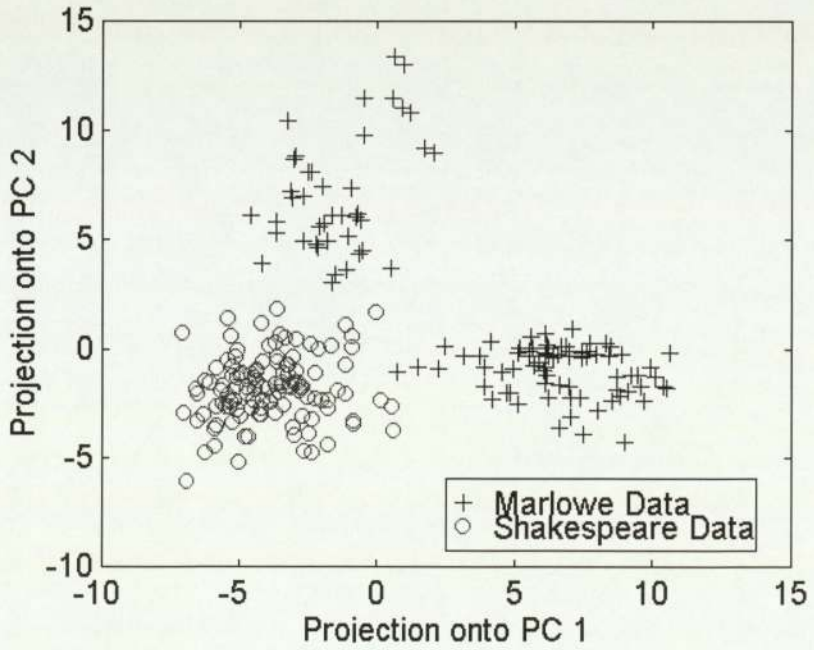

Figure 3.8 Projection of the 473 digram data set onto the first two principal components using a 3000 digram sample size

It can be seen that the problem almost appears to be of a linear nature based on just the first two principal component projections. It would be expected that a linear network classifier should produce very good results. Figure 3.9 shows a plot of the singular values corresponding to the first thirty principal components. After the fifth value, the values no longer decline smoothly indicating that the majority of the underlying variance in the data due to author differences has again been described by the first five principal components. The remaining principal components are mainly describing the noise in the data.

The projections of the training set onto the first five principal components were used to train a 5-input linear network classifier. The results of applying the network to a validation set generated from the validation data plays are illustrated in Table 3.17 and Table 3.18.

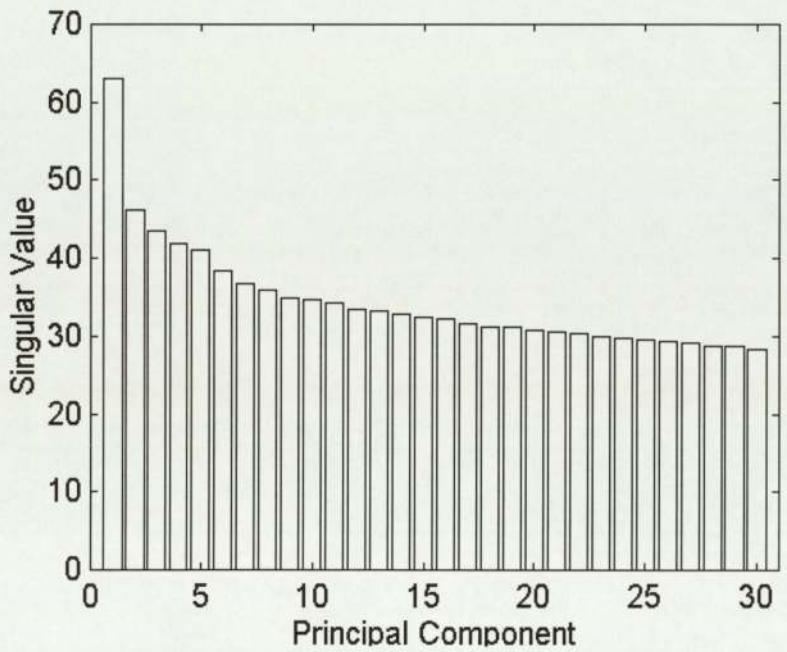

Figure 3.9 First 30 Singular Values obtained when PCA is performed on the 473 digram data set

	Predicted as Shakespeare	Predicted as Marlowe
Actual Shakespeare	29	1
Actual Marlowe	0	30

Table 3.17 Confusion Matrix obtained using Linear Network trained using core canon Shakespeare and Marlowe samples with 5 principal component projections as inputs on validation set

P(Shakespeare | Shakespeare Predicted)	**1.0**
P(Shakespeare | Marlowe Predicted)	**0.03**
P(Marlowe | Marlowe Predicted)	**0.97**
P(Marlowe | Shakespeare Predicted)	**0.0**

Table 3.18 Posterior probabilities derived from the results of applying a Linear Network trained using core canon Shakespeare and Marlowe samples with 5 principal component projections as inputs to validation set

As expected, the linear network produces excellent discrimination results, misclassifying only one sample. However, a problem arises when the network is applied to a new test set generated from plays that are not in the core canons of either author. This new test set contains 30 Marlowe and 30 Shakespeare samples generated from the plays listed in Table 3.19.

Marlowe Test Data	Shakespeare Test Data
Massacre at Paris (30527)	Troilus and Cressida (78912)
Dido (42425)	King Lear (77688)
Doctor Faustus (36200)	The Tempest (48779)

Table 3.19 List of non core canon Marlowe and Shakespeare plays used for test data

The result of applying the linear network to this new test set is illustrated in Table 3.20. Performance remains high on the non core canon Shakespeare material. However, this is not the case with samples from the non core canon Marlowe plays. This is probably due to the questionable nature of these plays as identified in Section 1.2.

	Predicted as Shakespeare	**Predicted as Marlowe**
Actual Shakespeare	29	1
Actual Marlowe	15	15

Table 3.20 Confusion Matrix obtained using Linear Network trained using core canon Shakespeare and Marlowe samples with 5 principal component projections as inputs on test set

By using a large enough sample size, a data set can be constructed containing the frequencies of the 473 digrams over each entire play. Table 3.21 shows the result of applying the same linear network to each play in the validation and test sets as well as to the three disputed plays. Lowe & Matthews (1995) introduce a simple method

of interpreting the outputs of any network based classifier model. A slight adaptation of this technique will be used. During training, each Marlowe sample is presented to the network with the ordered pair [1,0] as a target output. Conversely, each Shakespeare sample is presented with a target output pair of [0,1]. Hence, ideally any previously unseen Marlowe sample when presented to the trained network would produce the output pair [1,0]. A Characteristic Marlowe Indicator (CMI) value can be used to illustrate how close to 'Marlowe-like' the actual network output is. The CMI value is calculated using the following formula

$$CMI = (\mathbf{O} - \mathbf{T_S})^2 / ((\mathbf{O} - \mathbf{T_S})^2 + (\mathbf{O} - \mathbf{T_M})^2) \qquad \textbf{(3.3)}$$

where **O** is the network output, $\mathbf{T_M}$ is the target output pair for a Marlowe play and $\mathbf{T_S}$ is the target output pair for a Shakespeare play. A CMI value of '1' indicates full Marlowe characteristics and a value of '0' indicates full Shakespeare characteristics.

Author	Play	CMI	Prediction
Marlowe	Tamburlaine I	0.9985	Marlowe
Marlowe	Tamburlaine II	0.9932	Marlowe
Marlowe	Edward II	0.9759	Marlowe
Shakespeare	Comedy of Errors	0.0069	Shakespeare
Shakespeare	A Midsummer Night's Dream	0.0825	Shakespeare
Shakespeare	All's Well That Ends Well	0.0161	Shakespeare
Marlowe	Massacre at Paris	0.6715	Marlowe
Marlowe	Dido	0.5907	Marlowe
Marlowe	Doctor Faustus	0.1783	**Shakespeare**
Shakespeare	Troilus and Cressida	0.1174	Shakespeare
Shakespeare	King Lear	0.0629	Shakespeare
Shakespeare	The Tempest	0.1341	Shakespeare
Disputed	Henry VI part 2	0.5214	**Marlowe**
Disputed	Henry VI part 3	0.8053	**Marlowe**
Disputed	Titus Andronicus	0.3107	**Shakespeare**

Table 3.21 Linear Network Predictions for entire plays in validation, test and disputed sets. Network was trained using core canon Shakespeare and Marlowe samples

The CMI values for the plays *Massacre at Paris*, *Dido* and *Doctor Faustus* support the argument that these plays may contain external non-Marlovian material. Only *Massacre at Paris* could be argued to have a high enough CMI to suggest true Marlowe authorship. Indeed, *Doctor Faustus* is classified by the network to have been

written by Shakespeare. It is interesting to note that the network actually classifies the second and third parts of *Henry VI* as being characteristic of Marlowe with the third part having very strong Marlovian characteristics. *Titus Andronicus* on the other hand, is classified to be more characteristic of Shakespeare.

An MLP classifier model was also constructed which used the same principal component projections that were used to train the linear network classifier model. Unsurprisingly, this produced no advantages over the linear network, producing identical confusion matrices when applied to both the validation and test data sets.

On closer examination of Figure 3.8 it can be seen that the principal component projections for the three Marlowe training plays are divided into two distinct groupings. The smaller grouping of Marlowe projections corresponds solely to samples from *Edward II*. The larger grouping consists of principal component projections produced from samples of *Tamburlaine I* and *Tamburlaine II*. The distance between the two groupings, and therefore the apparent difference between plays written by the same author, does suggest that three is an insufficient number of Marlowe plays to use in training a classifier model.

A method which might lead to an improvement in the classification of the non core canon Marlowe data would be to expand the number of Marlowe plays from which the training data is generated. Although these additional plays may contain some non Marlowe material, it would be interesting to see what effect their inclusion might have.

A new training set was constructed containing 81 Marlowe samples and 81 Shakespeare samples generated from the plays listed in Table 3.1 together with the two plays *Massacre at Paris* and *Doctor Faustus*. Once again, a sample size of 3000 digrams was used. PCA was performed on the training set and all of the data sets were replaced by the projections onto the first five principal components as before. A new test set was also generated using the plays listed in Table 3.22 as samples from two of the plays in the original test set are now being used during training.

Figure 3.10 illustrates the projection of the play samples in the new training data set onto the first two principal components. It can be seen that although there still appears to be a good initial discrimination between samples from both authors, there is now some encroachment of Marlowe projections into the Shakespeare

Marlowe Test Data	Shakespeare Test Data
Dido (42425)	Troilus and Cressida (78912)
Hero and Leander (21172)	King Lear (77688)
The Jew of Malta (54282)	The Tempest (48779)

Table 3.22 List of non core canon Marlowe and Shakespeare plays used for new test data

grouping and vice versa. However, differences between individual plays written by Marlowe are now no longer as evident as those in Figure 3.8.

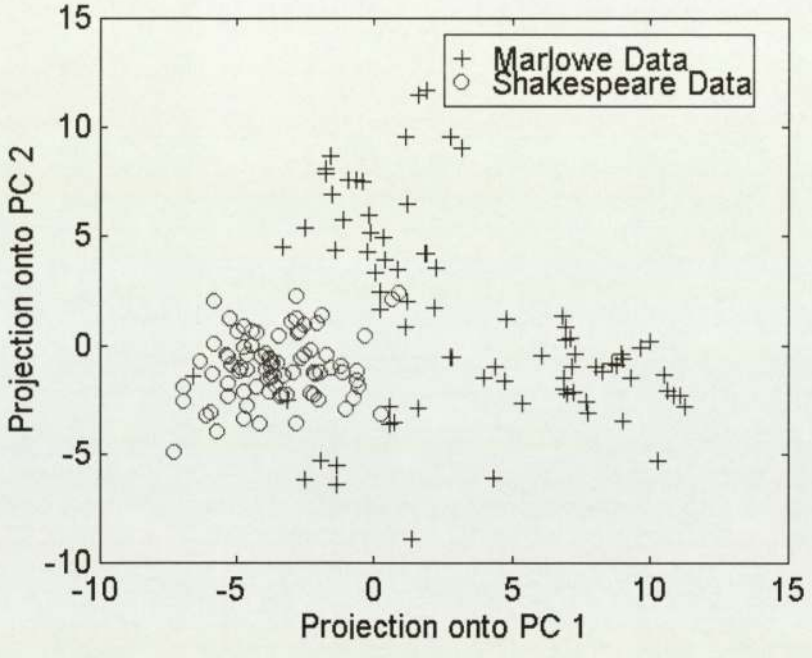

Figure 3.10 Projection of the 473 digram data set onto the first two principal components using a 3000 digram sample size and samples from core canon Shakespeare and both core canon and non core canon Marlowe plays

A linear network classifier model was trained on the new training data set and the result of testing the network on the original validation set is illustrated as a confusion matrix in Table 3.23. The associated posterior probabilities are given in Table 3.24.

Table 3.25 illustrates the result of applying the network to the new test set. Performance is degraded slightly when classifying the non core canon Shakespeare data. However an improvement is obvious with the classification of the questionable Marlowe samples.

	Predicted as Shakespeare	Predicted as Marlowe
Actual Shakespeare	27	3
Actual Marlowe	0	30

Table 3.23 Confusion Matrix obtained using Linear Network trained using core canon Shakespeare and both core canon and non core canon Marlowe samples with 5 principal component projections as inputs on validation set

P(Shakespeare | Shakespeare Predicted)	**1.0**
P(Shakespeare | Marlowe Predicted)	**0.09**
P(Marlowe | Marlowe Predicted)	**0.91**
P(Marlowe | Shakespeare Predicted)	**0.0**

Table 3.24 Posterior probabilities derived from the results of applying a Linear Network trained using core canon Shakespeare and both core canon and non core canon Marlowe samples with 5 principal component projections as inputs to validation set

	Predicted as Shakespeare	Predicted as Marlowe
Actual Shakespeare	24	6
Actual Marlowe	11	19

Table 3.25 Confusion Matrix obtained using Linear Network trained using core canon Shakespeare and both core canon and non core canon Marlowe samples with 5 principal component projections as inputs on test set

Once again, the results in Table 3.23 indicate that whilst a Linear Network produces good classification on the validation set, the problem does not appear to be completely linearly separable. This is supported by the confusion matrix for the test set given in Table 3.25. A neural network classifier may produce better results. An MLP Network with 5 principal component projection inputs and 2 hidden units was trained on the core canon Shakespeare and both the core canon and non core canon Marlowe samples. Table 3.26 and Table 3.27 illustrate the results of applying the trained network to the samples contained in the validation set.

	Predicted as Shakespeare	Predicted as Marlowe
Actual Shakespeare	28	2
Actual Marlowe	0	30

Table 3.26 Confusion Matrix obtained using an MLP Network trained using core canon Shakespeare and both core canon and non core canon Marlowe samples with 5 principal component projections as inputs on validation set

P(Shakespeare | Shakespeare Predicted)	**1.0**
P(Shakespeare | Marlowe Predicted)	**0.06**
P(Marlowe | Marlowe Predicted)	**0.94**
P(Marlowe | Shakespeare Predicted)	**0.0**

Table 3.27 Posterior probabilities derived from the results of applying an MLP Network trained using core canon Shakespeare and both core canon and non core canon Marlowe samples with 5 principal component projections as inputs to validation set

Two hidden units were once again chosen by training MLP networks of increasing complexities to completion and plotting the final sum-of-squares error over the training and validation sets. Figure 3.11 and Figure 3.12 as before illustrate the average final sum-of-squares error using five different MLP networks for each hidden unit count. Figure 3.12 shows that generalisation performance deteriorates when more than 2 hidden units are used.

Table 3.28 gives the results of applying the trained MLP network to the samples generated from the plays in Table 3.22. By comparing this table to Table 3.25, an obvious improvement is evident when using the neural network based classifier suggesting that some non-linearity is indeed present.

	Predicted as Shakespeare	**Predicted as Marlowe**
Actual Shakespeare	26	4
Actual Marlowe	7	23

Table 3.28 Confusion Matrix obtained using an MLP Network trained using core canon Shakespeare and both core canon and non core canon Marlowe samples with 5 principal component projections as inputs on test set

The MLP Network can also be applied to the Principal Component projections of the frequencies of the digrams over the entire plays. The results of this are illustrated in Table 3.29.

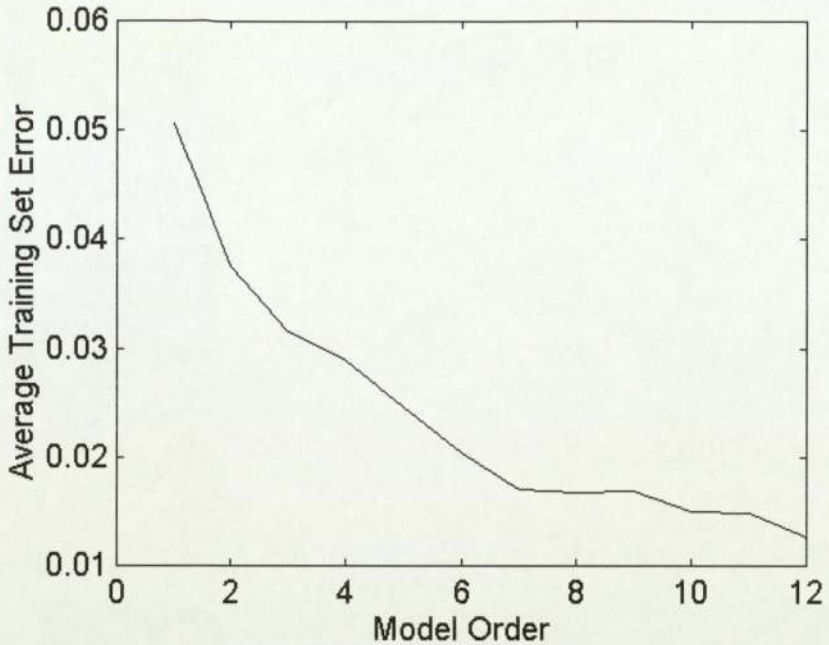

Figure 3.11 Training Set Error obtained after training MLP networks with different numbers of hidden units using core canon Shakespeare and both core canon and non core canon Marlowe samples

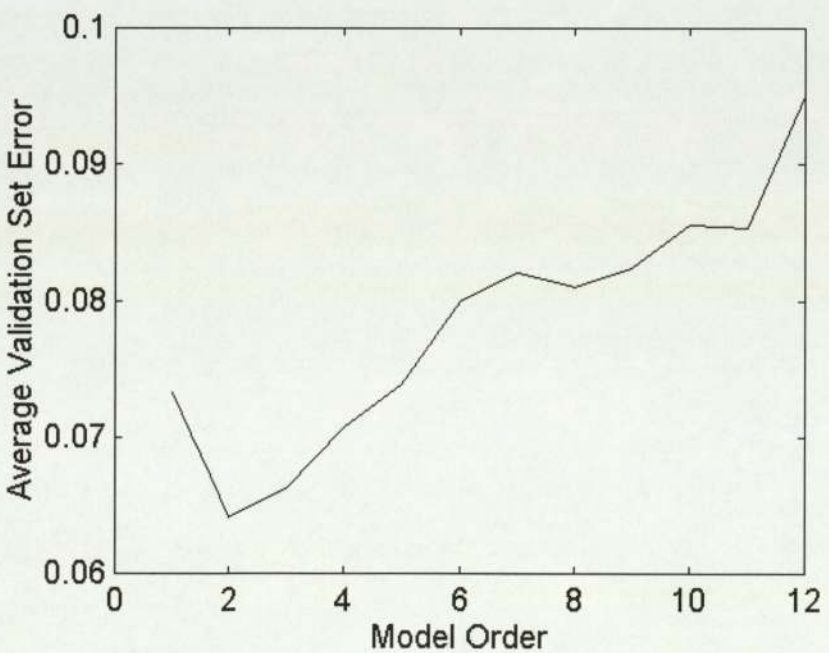

Figure 3.12 Validation Set Error obtained after training MLP networks with different numbers of hidden units using core canon Shakespeare and both core canon and non core canon Marlowe samples

Author	Play	CMI	Prediction
Marlowe	Tamburlaine I	0.9985	Marlowe
Marlowe	Tamburlaine II	0.9948	Marlowe
Marlowe	Edward II	0.9749	Marlowe
Shakespeare	Comedy of Errors	0.0437	Shakespeare
Shakespeare	A Midsummer Night's Dream	0.1531	Shakespeare
Shakespeare	All's Well That Ends Well	0.0523	Shakespeare
Marlowe	Dido	0.6732	Marlowe
Marlowe	Hero and Leander	0.8157	Marlowe
Marlowe	The Jew of Malta	0.4051	**Shakespeare**
Shakespeare	Troilus and Cressida	0.2111	Shakespeare
Shakespeare	King Lear	0.1328	Shakespeare
Shakespeare	The Tempest	0.2210	Shakespeare
Disputed	Henry VI part 2	0.7138	**Marlowe**
Disputed	Henry VI part 3	0.8804	**Marlowe**
Disputed	Titus Andronicus	0.6024	**Marlowe**

Table 3.29 MLP Network Predictions for entire works in validation, test and disputed sets. The Network was trained using samples from core canon Shakespeare and both core canon and non core canon Marlowe plays

The advantage of RBF networks over MLP networks for problems involving low amounts of training data has been identified in Section 2.3.2. Since the authorship dispute concerning Shakespeare and Marlowe falls into this category, the use of an RBF network may prove beneficial.

The appropriate number of basis function centres is chosen using the same method that is used to select the number of hidden units for an MLP network. RBF networks of increasing complexities were trained to completion and the final sum-of-squares error over the training and validation sets were plotted. Figure 3.13 and Figure 3.14 illustrate the average final sum-of-squares errors using twenty different RBF networks for each hidden unit count. The minimum of the curve in Figure 3.14 corresponds to 16 hidden units and this was the number of basis function centres chosen for the final network. However, any value between 12 and 19 could probably be used as there is such a small difference in the final validation errors corresponding to basis function counts in this range.

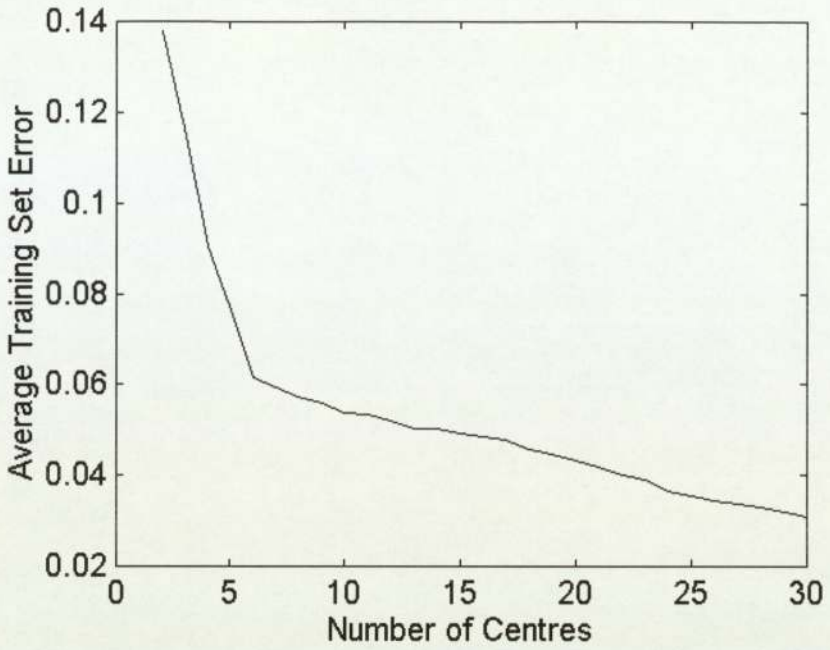

Figure 3.13 Training Set Error obtained after training RBF networks with different numbers of basis function centres using core canon Shakespeare and both core canon and non core canon Marlowe samples

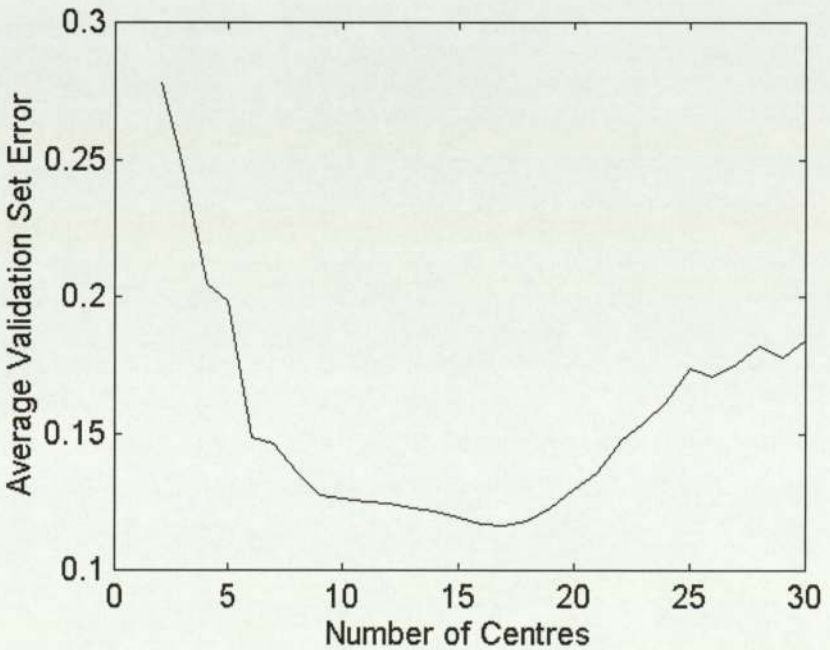

Figure 3.14 Validation Set Error obtained after training RBF networks with different numbers of basis function centres using core canon Shakespeare and both core canon and non core canon Marlowe samples

An RBF network with five principal component projection inputs and 16 hidden units was trained on the core canon Shakespeare and both the core canon and

non core canon Marlowe samples. Table 3.30 and Table 3.31 illustrate the results of applying the trained network to the samples contained in the validation set.

	Predicted as Shakespeare	Predicted as Marlowe
Actual Shakespeare	28	2
Actual Marlowe	1	29

Table 3.30 Confusion Matrix obtained using an RBF Network trained using core canon Shakespeare and both core canon and non core canon Marlowe samples with 5 principal component projections as inputs on validation set

P(Shakespeare | Shakespeare Predicted)	**0.97**
P(Shakespeare | Marlowe Predicted)	**0.06**
P(Marlowe | Marlowe Predicted)	**0.94**
P(Marlowe | Shakespeare Predicted)	**0.03**

Table 3.31 Posterior probabilities derived from the results of applying an RBF Network trained using core canon Shakespeare and both core canon and non core canon Marlowe samples with 5 principal component projections as inputs to validation set

Table 3.32 illustrates the results obtained when the RBF network is applied to the samples contained in the test set. The results of applying the same network to the Principal Component projections of the frequencies of the digrams over the entire plays is given in Table 3.33.

	Predicted as Shakespeare	Predicted as Marlowe
Actual Shakespeare	26	4
Actual Marlowe	8	22

Table 3.32 Confusion Matrix obtained using an RBF Network trained using core canon Shakespeare and both core canon and non core canon Marlowe samples with 5 principal component projections as inputs on test set

Although its performance is clearly better than that of a linear network classifier, an RBF network appears to offer no advantage over an MLP network. In fact, the performance of the RBF network is actually less satisfactory with this problem than that of the MLP network used previously. An explanation of why RBF functions generally perform better with problems of low data availability is given in Section 2.3.2. It states that large MLP networks require a large data set to successfully be able to optimise every network parameter and that equivalent size

Author	Play	CMI	Prediction
Marlowe	Tamburlaine I	0.9992	Marlowe
Marlowe	Tamburlaine II	0.9949	Marlowe
Marlowe	Edward II	0.9959	Marlowe
Shakespeare	Comedy of Errors	0.0644	Shakespeare
Shakespeare	A Midsummer Night's Dream	0.1319	Shakespeare
Shakespeare	All's Well That Ends Well	0.0556	Shakespeare
Marlowe	Dido	0.7023	Marlowe
Marlowe	Hero and Leander	0.7969	Marlowe
Marlowe	The Jew of Malta	0.3158	**Shakespeare**
Shakespeare	Troilus and Cressida	0.2179	Shakespeare
Shakespeare	King Lear	0.1264	Shakespeare
Shakespeare	The Tempest	0.2255	Shakespeare
Disputed	Henry VI part 2	0.6452	**Marlowe**
Disputed	Henry VI part 3	0.8986	**Marlowe**
Disputed	Titus Andronicus	0.6100	**Marlowe**

Table 3.33 RBF Network Predictions for entire works in validation, test and disputed sets. The Network was trained using samples from core canon Shakespeare and both core canon and non core canon Marlowe plays

RBF networks have less parameters to optimise. However, it was discovered that an MLP network requires only two hidden units to model the data for this problem. With such a relatively small network, the total number of parameters is very small. The number is in fact comparable to the number of parameters in an RBF network that need to be optimised. Thus, the performance of both types of network may be expected to be very similar.

The two neural network models used produce very similar results when applied to digram frequencies over entire plays. These results are illustrated in Table 3.29 and Table 3.33. The core canon plays of both authors produce strong CMI values as would be expected. The motivation behind the use of some non core canon Marlowe material in the training data was to try to improve the performance on non core canon Marlowe plays. This appears to have happened. Both neural network models produce a higher, more favourable, CMI value for the only non core canon Marlowe play which appears in both test sets, *Dido*. The original value can be seen in Table 3.21. The figures for *The Jew of Malta* do seem to support the argument, outlined in Section 1.2, that the play has possibly been falsely credited to Marlowe. This play is also the cause of the largest difference in values produced by the two

neural network classifiers with the RBF network making it more 'Shakespeare-like' than the MLP network.

The inclusion of non core canon Marlowe material in the training data does however have an adverse effect on the classification of the non core canon Shakespeare plays. All three plays in the test set are assigned significantly higher CMI values by the two networks than the original network trained using only core canon material. The values assigned to *The Tempest* are particularly noteworthy, suggesting that the play is quite uncharacteristic of Shakespeare. This loss of performance is to be expected considering the additional noise on the training data.

Both types of network classify the second and third parts of *Henry VI* to have been produced by Marlowe as before. Again, the third part appears to be the most 'Marlowe-like'. However, the use of some non core canon Marlowe material in the training data actually leads to a change in the classification of the third of the three disputed plays, *Titus Andronicus*.

3.5 Comparison of Methods

A number of different approaches to the investigation of an authorship problem have been examined. The majority of methods are based on measures of the frequencies of certain digrams occurring throughout a text sample. The thesis actually follows the chronological development culminating with the final and apparently most successful technique which utilises principal component projections of a number of digram frequencies. It has been noted that with this method, once the digram frequencies are transferred to the new feature space, a non-linear model initially appears to offer little advantage over a linear model. The explanation for this is that a great deal of linear processing has already been performed by PCA. The data has already been reorganised in the feature space to describe the variance as well as possible. As it is probable that a great deal of this variance will correspond to some author characteristic differences, the data is positioned in the feature space so that patterns characteristic of one author tend to form a vague grouping and those of another form a separate grouping. This can be seen in two dimensions in Figure 3.8. A classifier presented with data in the feature space then has the relatively easy task of

identifying the two previously created groupings. In such circumstances a simple linear model should perform equally as well as a more complex model such as a neural network.

The various approaches used could all be combined to form a committee of networks. Linear, MLP and RBF networks using both digram frequencies and principal component projections of frequencies as discriminators could be applied to the same problem. The average of the outputs produced by each of the networks could be taken and used as the output of the committee. Committees often offer a better performance than individual networks as the errors on the outputs of separate component networks can be averaged out.

4. Sensitivity Issues

This section looks in more detail at the accuracy or reliability of the performance of a network classifier model. Two different techniques are examined. The first is based upon the relative positions of data samples in the feature space. The second is based upon a model of the error on the network outputs themselves.

4.1 Outliers

So far, no consideration has been given to the possibility that an author other than Shakespeare or Marlowe is responsible for a new sample presented to a classifier network. Consider the two dimensional 'toy' problem illustrated in Figure 4.1. A training data set consists of patterns belonging to two classes. Two clear groupings are evident when this data is transferred into a new feature space. It is clear that a new pattern corresponding to point A in the feature space should be classified as belonging to class 1. However, how should patterns corresponding to points B and C be classified?

The appearance of pattern C is fairly easy to interpret, it can be considered an 'outlier' to the original model of the two classes as it occurs so far from either of the class groupings. There could be little argument against stating that pattern C does not belong to either class in the model and instead belongs to a third class, class 3.

The pattern represented by point B falls in an area immediately between the two classes. This could indicate that although it belongs to one class, it is not totally characteristic of that class and has some characteristics of the other class. A second possibility is that it is a member of both classes at the same time. If the two class groupings represent the works of two authors then this situation might equate to the new pattern representing a sample taken from a collaborative work of both authors. A final possibility is that the new pattern belongs to a new class, which has

characteristics much closer to class 1 and 2 than class 3 which is responsible for pattern C.

This simple example can be extended for the authorship problem being investigated in this thesis. Instead of being two-dimensional the problem is N-dimensional where N is the number of inputs to a classifier model.

The Euclidean distance between a new pattern vector and a pattern vector representing the mean of a class grouping can be calculated using formula 2.3. A simple approach to deal with outliers in the data would be to eliminate all patterns whose distance from both class means is greater than some arbitrary threshold value. This approach is not feasible for the current problem due to large spatial differences within the Marlowe training samples. Both Figure 3.8 and Figure 3.10 demonstrate that even in two dimensions, there is a large distance between a high percentage of the Marlowe training data patterns and the position of the Marlowe pattern mean and this extends to the third, fourth and fifth dimensions. It would be very difficult to set a threshold value based on the Euclidean distance to the mean as a number of core canon Marlowe samples would probably fall outside any useful value.

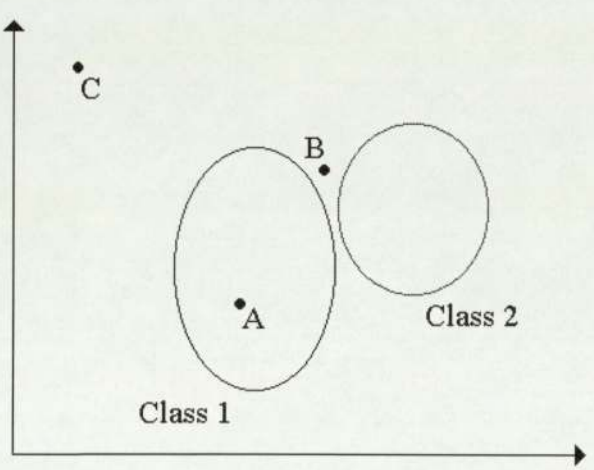

Figure 4.1 Two dimensional toy problem demonstrating problems of confidence in classification of a new pattern

An alternative approach would be to find the nearest neighbour pattern to the unclassified pattern. The Euclidean distance between the two vectors can be measured and used as an indication of how close the new pattern lies to the two modelled classes in the feature space. Recall that Table 3.21 illustrates the results of applying a linear network to the first five principal component projections of digrams over entire

play texts. It is important that only core canon plays are used during training. If samples taken from other plays are used, the distances between a new pattern and pattern vectors representing play samples produced by alternative authors may be calculated and used unintentionally. The results are repeated in Table 4.2. However, each network prediction is now based on both the CMI value and the distance between the vectors representing the play and its nearest neighbour in the training set. The predictions are produced using the following set of rules.

Evidence	Prediction
Very strong CMI	Author corresponding to CMI
High nearest neighbour distance and not very strong CMI	Alternative author
Low nearest neighbour distance and good CMI	Author corresponding to CMI
Low nearest neighbour distance and poor CMI	Characteristic of both authors

Table 4.1 Rules associating CMI values and nearest neighbour distances with network prediction

The appearance of an outlier with a large nearest neighbour distance combined with anything but a very strong CMI value (a value of 0 - 0.15 or 0.85 - 1.0, say) is interpreted to indicate alternative authorship. A play is classified to be characteristic of both authors if it has a relatively poor CMI value (a value of 0.4 - 0.6, say) but lies fairly close to at least one other sample vector. In all other cases, the play is predicted to have been produced by the appropriate author associated with the CMI value.

The predictions allocated to the non core canon Marlowe plays in Table 4.2 appear to support general scholarly belief. Both *Dido* and *Doctor Faustus* are classified to have the characteristics of an external author and these were suggested in Section 1.2 to contain a large amount of non Marlovian material. Only *Massacre at Paris* is predicted to have been written by Marlowe and this is believed to be the most Marlovian of the three non core canon plays. Previously the second and third parts of *Henry VI* were predicted to have been produced by Marlowe. Although this is still the case for *Henry VI part 3*, *Henry VI part 2* is now classified as having the characteristics of both authors. This does not necessarily mean that the play is a collaborative work of the two authors. It may be that the author of the play was influenced greatly by the style of the other author at the time of its production.

Author	Play	CMI	Nearest Neighbour Distance	Prediction
Marlowe	Tamburlaine I	0.9985	1.3722	Marlowe
Marlowe	Tamburlaine II	0.9932	1.9367	Marlowe
Marlowe	Edward II	0.9759	2.1389	Marlowe
Shakespeare	Comedy of Errors	0.0069	1.2567	Shakespeare
Shakespeare	A Midsummer Night's Dream	0.0825	1.7434	Shakespeare
Shakespeare	All's Well That Ends Well	0.0161	1.6841	Shakespeare
Marlowe	Massacre at Paris	0.6715	1.8559	Marlowe
Marlowe	Dido	0.5907	3.3387	**Other**
Marlowe	Doctor Faustus	0.1783	2.2465	**Other / Shakespeare**
Shakespeare	Troilus and Cressida	0.1174	1.8667	Shakespeare
Shakespeare	King Lear	0.0629	1.6819	Shakespeare
Shakespeare	The Tempest	0.1341	1.8939	Shakespeare
Disputed	Henry VI part 2	0.5214	1.6660	**Marlowe / Shakespeare**
Disputed	Henry VI part 3	0.8053	1.2949	**Marlowe**
Disputed	Titus Andronicus	0.3107	1.6506	**Shakespeare**

Table 4.2 Linear Network Predictions for entire plays in validation, test and disputed sets including nearest neighbour distances. Network was trained using core canon Shakespeare and Marlowe samples.

4.2 Error Bars

Error bars are a popular method of illustrating confidence in the output of a network classifier model. Lowe & Zapart (1997) suggest a number of methods for calculating error bar values. One approach is predictive error bar estimation. This technique will be used to produce error bars for the CMI values illustrated in Table 3.21. Although the use of neural network models are suggested, a linear network model can be used to generate the error bars. Whilst some performance may be lost to neural network error models, a simple approach is sensible in this case as the CMI values were calculated from the outputs of such a network.

With predictive error bar estimation, the network output produced by the original network for each sample in the training data set is used to calculate the local variance as follows

$$\text{variance} = \| \mathbf{t}_n - \mathbf{y}_n \|^2 \qquad \textbf{(4.1)}$$

where $\mathbf{t}_n$ is the target output vector and $\mathbf{y}_n$ is the actual output generated for input $\mathbf{x}_n$. The variances corresponding to each training data sample form a new target data set. A separate network is trained using the original inputs and these new targets. This models the squared error values, σ^2, on the original network outputs.

The error network, like the original classifier network, has two output nodes. On presentation of a new sample to the error network, the value of each output node represents the local confidence interval on the value of the corresponding output of the classifier network. As a linear network is used, the error predictions for both outputs are identical

The confidence interval provides a measure of the difference between the 'true' output pair which would be produced by a perfect model and the potentially erroneous output pair produced by the classifier network. The lowest possible 'true' value for one output of the classifier network can be calculated by subtracting the predicted error from the original output value. To compensate, this same error must also be added to the remaining output of the classifier network. The CMI value corresponding to this new output pair can then be calculated using formula 3.3. However, now the vector O' is substituted for O where O' is calculated using the following formula

$$O' = [(y_1 - \Sigma_1), (y_2 + \Sigma_2)] \qquad \textbf{(4.2)}$$

where the classifier network produces the output vector $[y_1, y_2]$ and the error network produces the output vector $[\Sigma_1, \Sigma_2]$. This CMI value describes a lower bound above which the 'true' CMI value must lie. Similarly, an upper bound can be implemented by substituting the vector O'' for O in formula 3.3 where

$$O'' = [(y_1 + \Sigma_1), (y_2 - \Sigma_2)] \qquad \textbf{(4.3)}$$

Figure 4.2 illustrates the CMI values assigned by a linear network to the plays in the validation, test and disputed data sets together with the calculated predictive

error bars on each value. The 'true' CMI value for each play must lie between the upper and lower bounds of the error bars in the diagram. A large error bar would suggest a relative lack of confidence in the network prediction. As expected, small error bars are associated with the works in the canons of both authors. The three non core canon Shakespeare plays also have relatively small error bars. However, this is not the case for both the non core canon Marlowe plays and the three disputed works.

Of all of the non core canon Marlowe plays, it might be expected that the prediction of *Doctor Faustus* as Shakespeare-like should have the largest associated error bars. Instead, the error bars associated with the CMI value for *Doctor Faustus* are the smallest of the three non core canon plays. This does not necessarily imply some confidence in the statement that *Doctor Faustus* was produced by Shakespeare. An alternative way of interpreting such a low CMI value combined with small error bars would be to state that some confidence can be placed in the suggestion that *Doctor Faustus* was not produced by Marlowe. It is suggested in Section 1.2 that *Doctor Faustus* contains a considerable amount of non Marlowe material added after the author's death. It may be that the style of whoever adapted the play before it was published was influenced by the style of an established William Shakespeare.

The large error bars associated with the CMI values for the remaining non core canon Marlowe plays do suggest considerable doubt in the predictions of the network. Again, this can be interpreted as support for the arguments in Section 1.2 that the majority of the remaining plays credited to Marlowe outside of the author's canon may not actually have been written by the author.

The CMI values for each of the three disputed plays also have large associated error bars. This would indicate that little confidence should be placed in the authorship prediction produced by the network for these plays. In fact, whereas the CMI value for *Henry VI part 2* is greater than 0.5, indicating Marlovian authorship, the lower bound actually lies at a value below 0.5 (Shakespearean authorship). These results do suggest that all three plays are not characteristic of a single author. Their closeness to the 0.5 CMI threshold value might also suggest that all contain the characteristics of both authors. However, caution must be taken with such suggestions as the same argument could be applied to the results for both *Massacre at Paris* and *Dido*.

Whilst the error bar estimation approach used above is a fairly simple one, the overall result does illustrate the problems associated with the accuracy of network predictions produced when using noisy data. It also demonstrates that the appearance of error bars can be interpreted in a number of different ways resulting in very different conclusions.

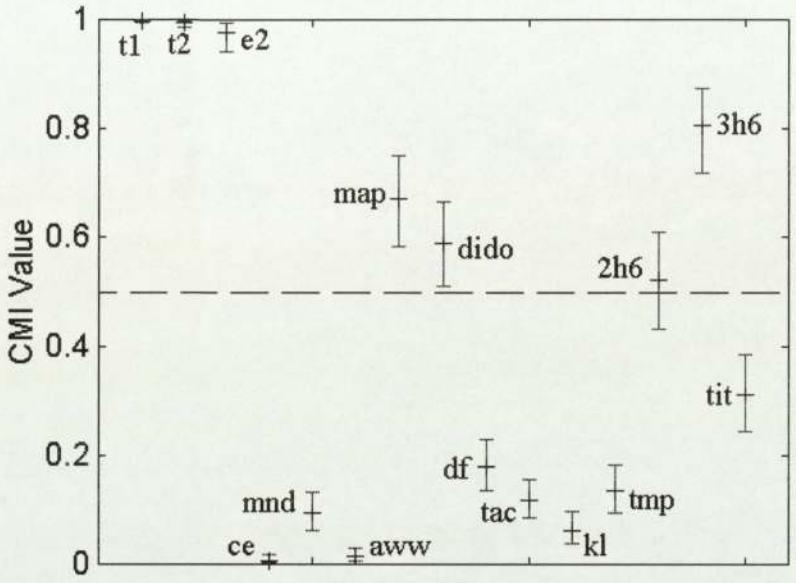

Figure 4.2 CMI values and corresponding error bars for entire plays, produced by training linear networks using core canon Shakespeare and Marlowe samples

5. Conclusions

One of the main conclusions of this thesis is the suitability of digram frequencies as stylometric discriminators to be used in the investigation of authorship disputes. A particularly noisy problem was chosen for investigation and the success of the new technique in this environment suggests that it should perform equally well when applied to other authorship problems.

The need for a lower amount of raw data necessary for training a classifier model also appears to have been proven. This was identified as a key requirement of the new technique in Section 1.4. A 3000 digram sample size corresponds to approximately 600 words. Matthews & Merriam (1994b) used the frequencies at which certain words occur over entire acts (3000 - 5000 words in length) as discriminators. A training set produced from a given number of plays using such a method will contain a significantly lower number of individual samples than one constructed using the technique outlined in this thesis.

A slight advantage has also been identified with the use of neural network techniques over linear techniques. However it is interesting to note that, in general, scholars using alternative discriminators such as word frequencies found a much greater improvement when using non-linear methods. This is probably due to the overall problem being reduced with the new technique to an, intermediary, almost linearly separable problem by projection into the principal component space as identified in Section 3.5.

The greatest potential for a classifier network based on digram frequencies would probably be as a component part of a network committee. Networks using different types of discriminators as inputs may be able to detect different author characteristic patterns in the data. Their knowledge can then be combined to produce a more educated prediction as to the true authorship of a given text.

All of the approaches used in this thesis do seem to conclude that the three disputed plays contain characteristics of both Christopher Marlowe and William Shakespeare, with the third part of *Henry VI* being the most 'Marlowe-like'. This agrees with the conclusions of other scholars working in this area. It would also lend support to the so called 'Marlovian Theory' surrounding *Henry VI parts 2 and 3* which suggests that the two plays were Shakespearean adaptations of original Marlowe texts.

Results throughout the thesis would also suggest that the plays credited to Marlowe outside of the author's canon have been greatly affected by noise. This may be external noise caused by adaptations to the author's works at a later date as proposed in Section 1.2. It may also be internal noise caused by the author's style varying considerably over his career. Alternatively the results might support claims that Marlowe was not actually responsible for a number of the texts generally credited to him.

There is scope for future work in this area. In this thesis only simple methods of expressing confidence in a network's outputs have been investigated. These have, however, demonstrated the benefits of such measures in the final interpretation of the results. The utilisation of more sophisticated error bar estimation techniques might be advantageous, particularly in the investigation of the true authorship of some of the plays believed to be falsely credited to Marlowe.

It is worth noting that any set of network outputs can be interpreted in a number of different ways depending upon the hypothesis being investigated. For this reason, the network results in this thesis can only be used to either support or contradict arguments which have previously been suggested and not to formulate them.

References

Alt, M. (1990), *Exploring Hyperspace, A non-mathematical explanation of multivariate analysis*, Maidenhead: McGraw-Hill.

Bailey, R.W. (1979), "Authorship attribution in a forensic setting", *Proceedings of the Fifth International Symposium on Computers in Literary and Linguistic Research*, Birmingham: AMLC, pp. 1-20.

Bishop, C.M. (1995), *Neural Networks for Pattern Recognition*, Oxford: Clarendon Press.

Campbell, M.J. & Machin, D. (1993), *Medical Statistics A Commonsense Approach*, Chichester: Wiley.

Dear, I.C.B. (1986), *Oxford English A Guide to the Language*, Oxford: Oxford University Press.

Haykin, S.S. (1994), *Neural Networks, a comprehensive foundation*, Maxwell Macmillan.

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

Holmes, D.I. (1994), "Authorship Attribution", *Computers and the Humanities*, 28, 2, 87-106.

Lowe, D. & Matthews, R.A.J. (1995), "Shakespeare Vs. Fletcher: A Stylometric Analysis by Radial Basis Functions", *Computers and the Humanities*, 29, 449-461.

Lowe, D. & Zapart, K. (1997), "Validation of Neural Networks in Automotive Engine Calibration", *Proceedings of the Fifth IEE International Conference on Artificial Neural Networks*, Cambridge, pp. 221-226.

Ledger, G. & Merriam, T. (1994), "Shakespeare, Fletcher and the Two Noble Kinsmen", *Literary and Linguistic Computing*, 9, 3, 235-248.

Matthews, R.A.J. & Merriam, T. (1994a), "A Bard by any other name", *New Scientist*, 22 January, 23-27.

Matthews, R.A.J. & Merriam, T. (1994b), "Using Neural Networks to Cast Light on Literary Mysteries", *Proceedings of Expert Systems '94, 14th Annual Conference of British Computer Society Specialist Group on Expert Systems*, Cambridge, 2, pp. 237-247.

Merriam, T. (1995), "Possible Light on a Kyd Canon", *Notes and Queries*, 240:3, 340-341.

Merriam, T. (1996), "Tamburlaine Stalks in Henry VI", *Computers and the Humanities*, 30, 267-280.

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

Myers, W. (1990), "Somebody Wrote Shakespeare - The Computer as Literary Detective, Part 2", *IEEE Expert*, 5, 4, 84-85.

Tweedie, F.J., Singh, S. & Holmes, D.I. (1996), "Neural Network Applications in Stylometry - The Federalist-Papers", *Computers and the Humanities*, 30, 1, 1-10.

Udny Yule, G. (1968), *The Statistical Study of Literary Vocabulary*, Hamden: Archon Books.

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

A. Investigation of Optimum Sample Size

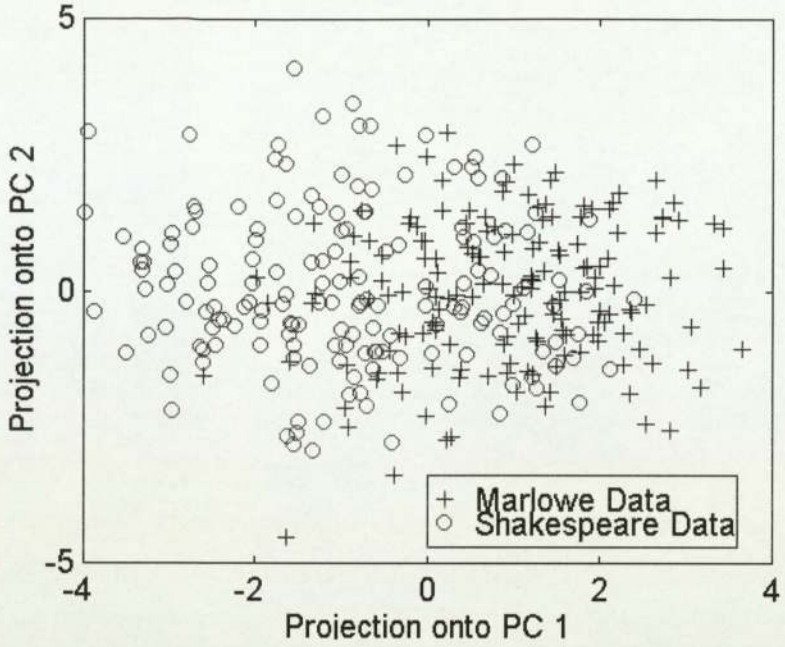

Figure A.1 Projection of the 25 digram data set onto the first two principal components using a 1000 digram sample size

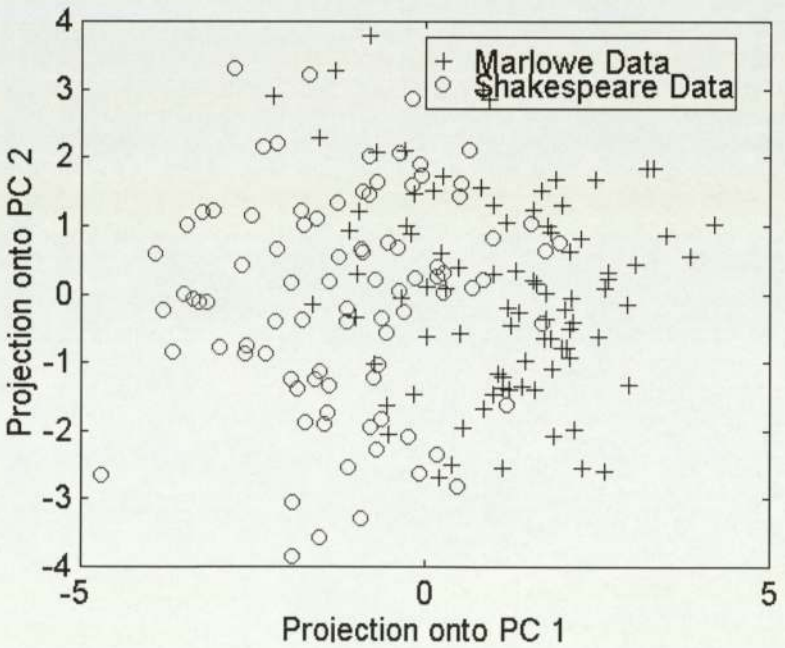

Figure A.2 Projection of the 25 digram data set onto the first two principal components using a 2000 digram sample size

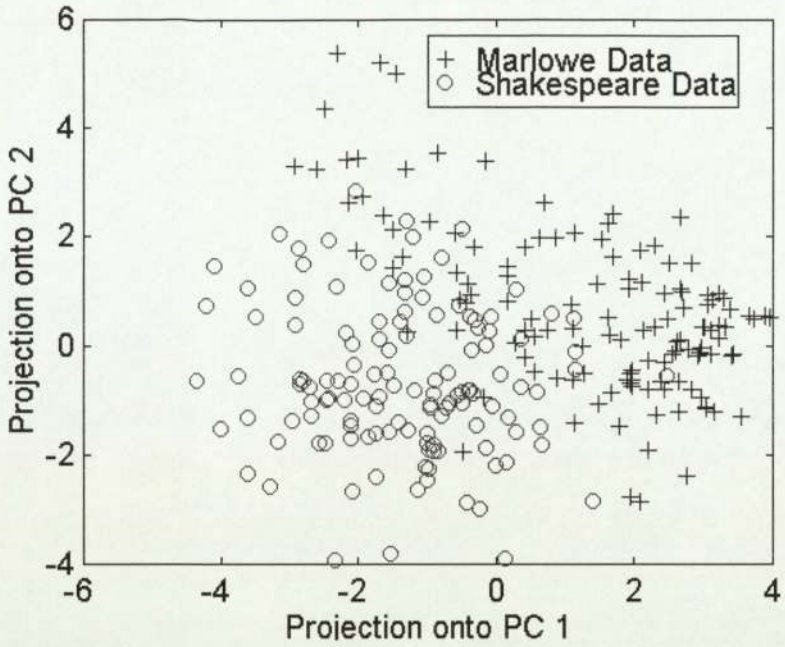

Figure A.3 Projection of the 25 digram data set onto the first two principal components using a 4000 digram sample size

B. Determination of Digrams which contribute most to PCA

Digram	Eigenvector 1	Eigenvector 2	Eigenvector 3	Eigenvector 4	Eigenvector 5
AN	**0.1830**	**-0.1628**	**0.3073**	**-0.2329**	**-0.3952**
AR	-0.0579	-0.0813	0.0244	-0.1788	-0.3137
AT	-0.0436	0.1630	0.0000	-0.0280	0.0598
EA	0.0552	-0.1342	-0.1461	0.0260	-0.0564
EN	0.0232	-0.1598	-0.0089	0.1148	-0.0696
ER	-0.0542	0.2754	-0.0938	0.3009	-0.2011
ES	0.0432	-0.1808	-0.2353	0.0028	0.0849
HA	**-0.2164**	**0.3171**	**0.0592**	**-0.3015**	**-0.0348**
HE	**0.1542**	**0.4841**	**-0.2350**	**0.3198**	**-0.5278**
HI	-0.1358	-0.0329	-0.3264	-0.1546	-0.0693
IN	-0.1864	-0.0303	-0.3517	-0.1472	0.0977
IS	**0.2505**	**-0.0149**	**0.1387**	**0.3328**	**0.3567**
IT	-0.0820	0.1023	0.1287	0.0393	0.0420
LL	-0.0973	0.1361	0.0569	-0.2548	-0.1288
ME	-0.1495	0.0204	-0.0125	-0.0989	-0.0881
ND	**0.2812**	**-0.2854**	**0.2535**	**-0.2058**	**-0.3038**
NO	**-0.3200**	**0.1080**	**-0.0682**	**-0.0952**	**0.0661**
ON	-0.1328	-0.0067	-0.1175	-0.1216	-0.0292
OR	0.1168	-0.0746	-0.2455	-0.2776	0.0550
OU	**-0.5563**	**0.0729**	**0.4949**	**0.0319**	**0.0489**
RE	0.0031	0.0071	-0.1723	-0.1546	-0.0287
ST	0.0340	-0.0791	-0.1821	-0.1878	0.1654
TH	**0.4355**	**0.5573**	**0.1421**	**-0.4101**	**0.2784**
TO	-0.1074	-0.0227	-0.1189	-0.0938	-0.1652
VE	-0.0573	0.0167	-0.0780	0.0483	0.0533

Table B.1 Values of the loadings on the first five eigenvectors produced by PCA on a data set containing the frequencies of the 25 most common digrams

Single Letter	Eigenvector 1	Eigenvector 2	Eigenvector 3
A	0.1987	-0.1948	0.1019
B	-0.1380	0.0547	0.2050
C	0.0467	-0.2464	0.1206
D	0.1424	-0.1063	-0.2468
E	**0.1331**	**0.4000**	**-0.2271**
F	**0.1760**	**-0.2536**	**0.2119**
G	0.0010	0.0219	0.0047
H	**0.0725**	**0.3652**	**0.3486**
I	-0.1255	0.1180	0.2232
J	0.1194	-0.0543	0.1735
K	**-0.2574**	**0.2106**	**-0.0323**
L	-0.0816	0.1234	-0.2250
M	-0.0637	0.1114	-0.3481
N	-0.0156	-0.1118	0.1026
O	**-0.4248**	**-0.2140**	**0.0007**
P	0.2080	-0.0050	-0.0570
Q	0.1163	-0.1254	-0.1246
R	**0.2780**	**-0.1285**	**-0.3017**
S	**0.3462**	**-0.0398**	**-0.0280**
T	0.0156	0.1822	0.4163
U	**-0.2359**	**-0.4575**	**0.0840**
V	0.0846	0.0759	-0.1647
W	**-0.2804**	**0.2168**	**-0.1428**
X	-0.0154	0.1468	0.1350
Y	**-0.3157**	**-0.1746**	**-0.1682**
Z	0.2859	-0.0219	0.0759

Table B.2 Values of the loadings on the first three eigenvectors produced by PCA on a data set containing the frequencies of the 26 single letters