

A blended text mining method for authorship authentication analysis

Philip Sallis and Subana Shanmuganathan

Auckland University of Technology

philip.sallis@aut.ac.nz; subana.shanmuganathan@aut.ac.nz

Abstract

The paper elaborates upon the interim results achieved in resolving a few newly discovered 16th century letters now alleged to be written by Queen Mary of Scots (QMS). Despite the significant progress seen in stylometry and its role in authorship attribute analysis especially in disputed writings/ texts controversies over the authorship of Shakespeare's literary work still continue as does research into this corpus of letters. Using more sophisticated computational and mathematical modelling techniques than in previously published research, this study still employs the use of stylometric measures, to show a distinct variation between the authentic writings of QMS and the newly discovered letters, claimed by numerous enthusiasts to be of her authorship. Incorporating additional advanced statistical methods, such as principle component analysis (PCA) and artificial neural networks (ANNs), especially Kohonen's self-organising map (SOM) based visualisation technique, a text mining approach for this application has been developed. The similarities between different pairs of the new and authentic letters and in some cases within individual letters become apparent when using "cusum" analysis adding further complexity to the task of resolving the anomaly seen among QMS loyalists, archaeologists, linguists and the like. The reasons for the inconclusive results of this study are presented with suggestions for future work but in essence, the data mining method used is regarded as being unique in its blend of conventional and non-conventional statistics and useful for this class of text analysis problem

1. Introduction

The preliminary analysis carried out to resolve authorship authentication of newly discovered letters of 16th century writings lately alleged to have been written by Queen Mary of Scots (QMS) produced inconclusive results requiring further analysis. In order to overcome the preliminary analysis issues a blended text mining method has been developed. The paper gives a detailed report on the blended approach results following a brief introduction to stylometry and its role in authorship authentication of literary writings.

The origins of stylometry, the statistical analysis of literary style, could be traced back to the mid 1800s. Despite the significant development this area of study has undergone over the years there are not any precise procedures or methodologies to resolve a piece of disputed writing without any reproach (Holmes 1998). The situation leaves debates reopened over many literary writings. "... for every method that 'works' there soon appears counter arguments pointing out crucial flaws. A methodology successful for one attributional problem does not necessarily work for another ... " (Holmes 1998:1). The blended approach investigated herein uses conventional parametric statistics to analyse the writings and to describe similarities between different pairs of the new and authentic letters and in some cases mixed authorship within individual letters. The authorship issues become apparent with "cusum" analysis adding further complexity to the task of resolving the anomaly seen among QMS loyalists, archaeologists, linguists and the like. The various methods applied to resolving the anomalies relating to the disputed writings of QMS are discussed in detail along with areas proven to be best for the different methods chosen in this study. Finally, the reasons for the inconclusive results achieved through this study are presented with suggestions for future work. The significant aspect of this study is the data mining method used that could be regarded as being somewhat unique in its blend of conventional and non-conventional statistics and generally useful for analysing texts of this class with similar issues.

2. Stylometry in authorship attribute analysis of literary writings

Stylometry, the statistical analysis of literary style has its origins dating back to the mid 1800s. It is Augustus de Morgan, an English logician who first suggested to his friend in a letter in 1851 that "... questions of authorship might be settled by determining the length of words 'if one text does not deal in longer words than another'..." (Holmes 1998:1). Thomas Mendenhall in 1880 investigated this hypothesis painstakingly by measuring the lengths of several hundred thousands of words from the works of Bacon, Marlowe and Shakespeare and published his results subsequently. In

his results lately described to be legendary, it was shown that word length is not an effective authorial discriminator however Thomas found some similarities between Shakespeare and Marlowe. The latter finding is still being investigated with more sophisticated techniques using principal component analysis (PCA) and artificial neural networks (ANN), such as radial bias and multiple layer perceptrons (MLPs) as seen in (Merriam 1998; 2004; Lowe and Matthews 2005).

3. The blended approach and results

The blended approach investigated herein consists of conventional and SOM based collective analyses performed on stylometric and statistical features of the writings. The following are the details of the letters used in the analysis:

Letter 1, written to Elizabeth during Mary's one-year incarceration in Lochleven after her defeat at Carberry Hill. Mary truly believed that her cousin would come to her rescue but Elizabeth's friendliness was just a front. Letter 3: Mary's last letter to Elizabeth, stating her final requests as anticipated it would never be granted. Letter 4: Mary's last letter ever written in the early hours of the day of her execution.

Casket letter one: In Latin, French, Scots and English. The English version is in the Record Office. Casket letter two: Latin, French, Scots and English and stored in the Record Office. Casket letter three: There is no Latin or English version extant. A French version is in the Record Office State Papers and a Scots version appeared in Buchanan's "Detection". Casket letter four: French and English, at Hatfield. Also, Latin and Scots 'translations', and a published French version is available. Casket letter five: A French version is in the Record Office State Papers. It varies slightly from the published French version. There is also a version in Scots. A clerk has endorsed the Record Office version, "Anent the despatch of Margaret Carwood – which was before her marriage – proves her affection". Casket letter six: English and French at Hatfield. Also Scots and published and French. Casket letter seven: Scots and published French versions only.

Both stylometric¹ and statistical² features of Queen Mary's disputed letters are analysed individually and collectively using self-organising map (SOM) based techniques. Aaronson (1999), analysed the two features together (using k-means clustering) and produced better results than analysing them separately. Kohonen's SOM cluster analysis is successfully applied to visualising multidimensional data across a

¹ Stylometric; the statistical data drawn from parses of sentences.

² statistical ; the data driven statistics from the text itself.

wide range of disciplines. A SOM is single layered ANN that provides an excellent tool for visualising multidimensional datasets on low dimensional displays while preserving the details in the original dataset. Its application in this study provides a tool for visualising the complex data; the stylometric and other statistical features of QMS's authentic letters and disputed ones found in a casket to verify the authenticity of the latter.

SOM results

SOM results of both stylometric and statistical feature data (average) along with the readability scores³ (fog, flesch and kincaid) are discussed herein. Absolute value variables of stylometric features, such as number of characters (including and excluding blanks), word count and unique words are not included in the analysis. The variability of these variables depends on the length of the text analysed hence including them in this stylometric analysis for author attribution and authentication would be unconstructive to the analysis. Initially, a SOM (figures 1 a-f) is created with readability scores and its clustering consists of the following clusters (C1-C3):

C1: Casket letters one, two, four, six to eight along with the original letter 1 with ideal scores for readability; fog at 13.28 and flesch at 65.

C2: Casket letters three and five with the original letter 3 with not so ideal scores for readability; fog at 22.78 and flesch at 37.51.

C3: The original letters 2 and 4 have scores for fog 16.97 and flesch 49.46

Further clustering of C1 letters produced the following;

C1a: Casket letters one, four, and eight with the original letter 1 showing not so ideal scores for readability (fog at 14.5, flesch at 61.83095 and kincaid at 11.723625).

C1b: Casket letters two, six and seven in another cluster with the ideal scores (fog 11.655967, flesch 69.578233 and kincaid 9.1304667).

C2 letters are divided into two distinctive groups: C2 a: Casket letter three and the original letter 3 together and C2 b: Casket letter, five alone in another cluster.

C3 letters (original letters 2 and 4) are grouped into different clusters with insignificant difference between the two.

³ Readability scores fog, flesch and kincaid are calculated for text or block based on www.plainlanguage.com/Resources/readability.html
fog = words_per_sentence + percent_complex_words) * 0.4,
flesch = 206.835 - (1.015 * words_per_sentence) - (84.6 * syllables_per_word) and
kincaid = (11.8 * syllables_per_word) + (0.39 * words_per_sentence) - 15.59.

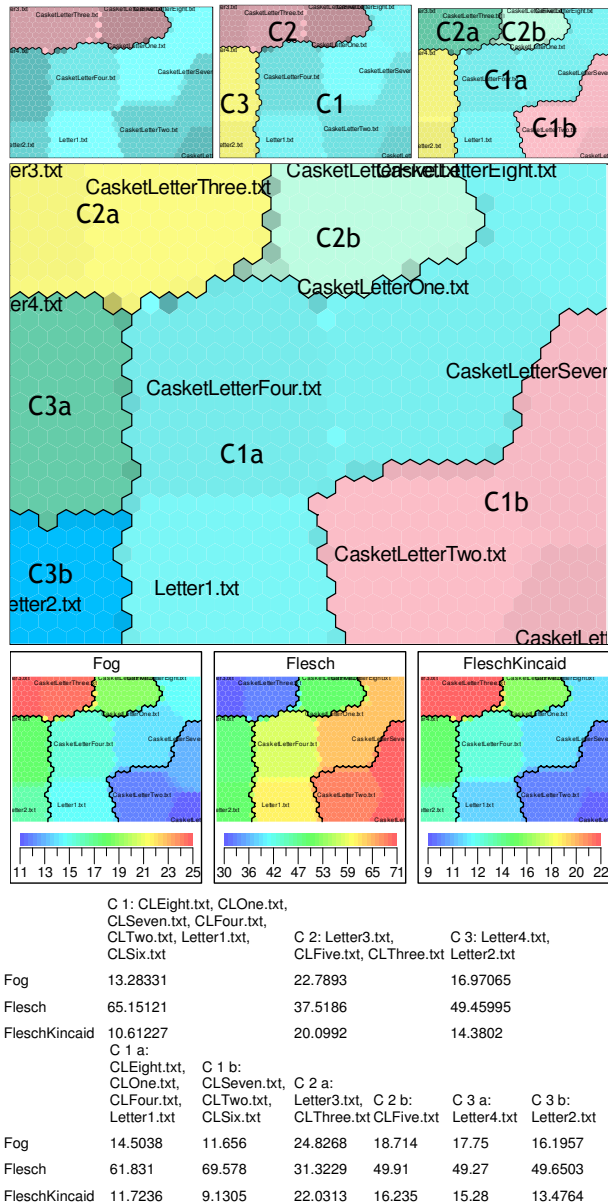


Figure 1 a-f: SOM (1000 nodes) of readability scores, a-d show the progressive SOM from 2 to 5, e: SOM component planes and f: Table showing the SOM cluster profiles.

In the next stage, all average values of stylistic features and readability scores are analysed together to see the clustering patterns in them and this should reveal useful information in the authentication of casket letters. The SOM (figures 2 a-e) created with all stylistic average count data and readability scores, consists of the following clusters:

C1: with casket letters one, two, four and six to eight show the highest readability scores with fairly low percent of complex words.

C2: with authentic letters 1 and 4.

C3: with casket letter three and the original letter 3 showing the highest average words per sentence.

C4: casket letter five with the lowest percentage for complex words.

C5: original letter 2 with the highest percentage for complex words.

Next, statistical data, such as percentages of word, sentence frequencies and percentage of letters generated using 'signature' software (Millican 2003) are analysed using SOMs leaving out the paragraph length as it considered being unconstructive in the analysis. The following are the SOM clustering details:

C2: Original letters 1, 2 and 4,

C3: Original letter 3 with casket letter 3 and C1a&b: Rest of the casket letters into one group.

Further clustering splits casket letters seven and eight from others (figure 3).

Function word analysis

The application of function word⁴ analysis to resolving authorship authentication of disputed texts produced promising results in (Binongo 2003; Farrington 2004). The use of different function words by an individual is described as a spontaneous and instantaneous act. It occurs in a subconscious state of human mind.

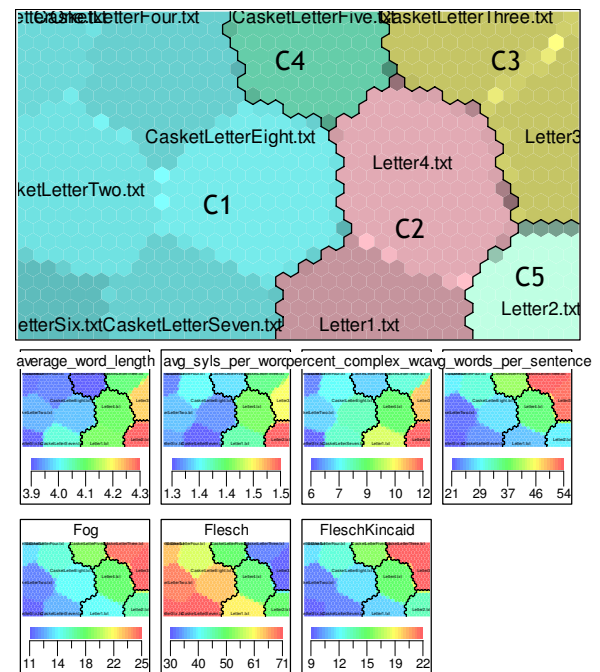


Figure 2 a: SOM of stylistic and statistical data of the original and casket letters of QMS and b: SOM components.

⁴ Function words or filler words, such as the, is, there, and, that and so on are used to join the noun, verb and similar main words to form a sentence.

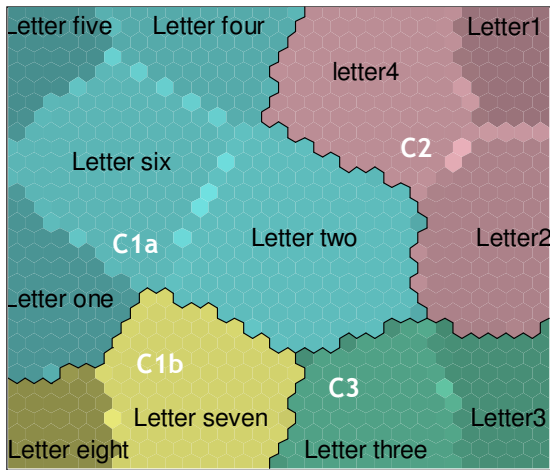


Figure 3: SOM of signature data (word length, letters and function word frequency) with readability scores, b: PCA of the same data.

Different individuals exhibit a characteristic way of using these words in their own writings and by studying the difference in the styles it is possible to prove/disprove the authorship of disputed texts. For example, Binongo (2003) showed the difference between Thompson's writings and Baum's original stories by analysing most frequently used 50 function words with PCA. The former took over the story writing after the death of the original author Baum, in an attempt to meet the overwhelming demand for the original writer of a popular children's fantasy series called "The Royal Book of Oz". Even though the publisher and Baum's widow stated that Thompson stories for the series were based on Baum's notes presumably to guarantee the sales, the percentages of the function word frequencies of the two authors clustered separately in the PCA. In consideration of this fact, frequency percentages of selected function words in all 12 letters are analysed using SOMs and then PCA herein. Seventeen function words found to be common in QMS's original letters 3 and 4 are used in the analysis. As far as the function word frequency percentages are concerned PCA and SOM analyses show similar clustering patterns in this analysis (figures 4 a & b):

Cluster 1: Casket letters 2, 4, 6, 7 and 8.

Cluster 2: Original letters 1, 2, 4 and casket letter 5.

Cluster 3: Original letter 3 and casket letters 1 and 3.

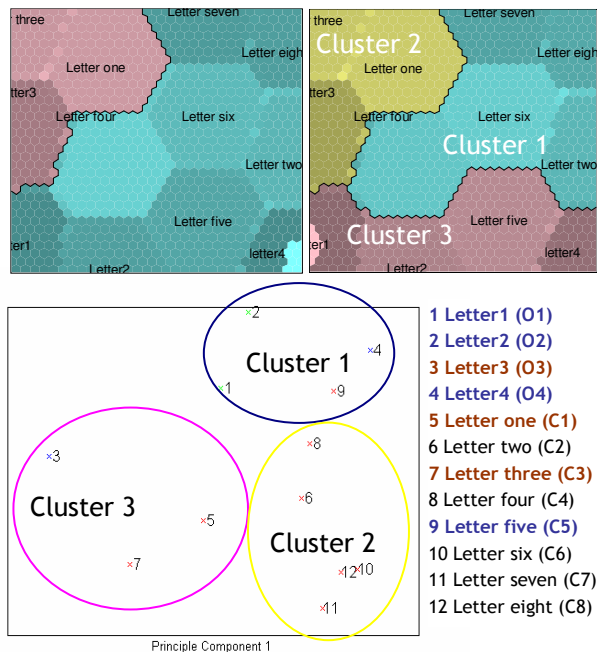
In this clustering, Queen Mary's authentic letters 3 and 4 are seen in different clusters. Possible reasons for this could be that the two letters were originally written in Scottish and French Languages respectively. Secondly, they were written to individuals with whom the author had different relationships. On the other hand, even though PCA of the function word

frequencies has been successful in differentiating texts of different authors, the approach is less reliable as far as different genre texts of a same author are concerned (Smith 2002), hence, might have clustered the original letters 3 and 4 separately. However, the efforts by Binongo (2003) that further subdivided the genre based clusters to distinguish the authors by analysing larger chunks of texts (10,000 words instead of the initial 5,000), this is not possible with Mary's letters as the letters are not so lengthy.

Cusum (Qsum) analysis

Finally, the letters are analysed using Cusum (qsum) methodology. This is an approach successfully applied to authorship attribution of short texts such as letters, or even spoken messages. In fact, the method is so renowned for its reliability it is even referred to as 'cusum fingerprints'. It is extensively applied to forensic authorship and attribution of short messages consisting of 25 to 50 sentences in a variety of text corpora, such as edited work, translation, children's utterance and anonymous works, and is explained herein based on (Buckland 1999; Farrington 2004).

The cusum method is a superimposition of two graphs: one on the sentence length distribution and the other on that of 'habit words'⁵ within the analysed text.



Figures 4 a & b: Two and three cluster SOM and PCA clustering of function word frequency percentages.

⁵ Habit words are the 2 to 3 letter function/ filler words and initial vowel words (ivw). Their use in texts and speeches is considered to be unique, reflecting an author's style (Farrington, J. M. (2004).

The sentence length distribution is the deviation in word length from the mean average sentence word length, calculated by finding the difference between the mean value and the number of words in each sentence. The deviation of habit words is also calculated again by finding the difference between the mean and real values of the use of these words in each and every sentence. In the end, the two sets of values are plotted on transparencies and superimposed to create a cusum chart that can be used to study the authorship attribution of the piece of writing. If the two lines track each other that would be an indication of single authorship and so the opposite implies mixed authorship of the writing.

The qsum chart of the original letters 3 and 4 (figure 5) deviate from the other letter; especially in the 3rd letter the graphs are far apart. In the 4th letter first two sentences deviate from the rest of the graph. This is something that could have resulted from the translation or the poignant state the author was in. Looking into the text, sentences 5-8 of the graph, illustrate the author's state, in which she makes requests as to what should be done to her remains when she is executed. In addition, as stated earlier, the two letters were originally written in Scottish and French, so possibly translated by different individuals, which is confirmed by qsum chart of the original letters 1-4.

Of the individual qsum charts of the original letters, 2 and 4 point to single authorship, whereas 1 and 3 point to either mixed authorship or some edition within them or the author's state of mind or the subject matter within these letters. Casket letters 1-8 show many discrepancies, the reasons for this being reflection of either translators' style or different genre. They also vary from the style of the original letters, demonstrated in the qsum chart of original letters 3 and 4 with casket letters three and five however, casket letter 1 and original letters 3 and 4, point to a signal authorship (figure 5a&b). There are too many reasons as to why qsum chart analysis is producing such contradicting results within the only 12 letters analysed herein.

4. Discussion

In the SOM analysis of readability scores alone, Queen Mary's original letters 2 and 4 were grouped together, whereas the original letter 3 was grouped with casket letters three and five. Rest of the casket letters and the original letter 1 were together with an ideal set of readability scores. However, Qsum chart analysis does not reflect this. SOM analysis of average data on word length, style word, words/sentence and percentage of complex words with readability scores grouped the

original letters 1, 2 and 4 with casket letter five. The original letter 3 was found grouped with casket letter three. Rest of the casket letters were grouped together. SOM created with percentages of word and letter frequencies (signature s/w) and readability grouped original letters 1, 2 and 4 together, original letter 3 with casket letter three and all the rest of the casket letters together.

Function word analyses carried out by SOM and PCA grouped the original letters 1, 2 and 4 with casket letter five. The original letter 3 was grouped with casket letter three and the rest of the casket letters were grouped into one. Analysis of function words are less reliable in authorship authentication and attribution as they tend to cluster texts based on their genre (Smith 2002). However, this could be overcome by increasing the text length as done in (Binongo 2003) where text chunks of 10,000 words were used instead of the initial 5,000 to zoom into the genre based clusters, and to subdivide them based on their authorship. This cannot be done in this study as Queen Mary's letters range from 100-3,600, not enough for this kind of analysis.

Based on the above stylometric and statistic analyses, casket letters 3 and 5 are more similar to Queen Mary's original letters. Apart from the normal aspects, such as subject matter, words learned over time and flow of the text (Aaronson 1999), one's emotional state of mind at the time of the writing as well could play a major role in the text produced by the individual concerned. Furthermore, qsum chart dose not confirm this either.

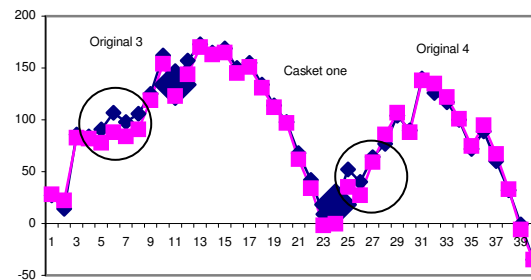
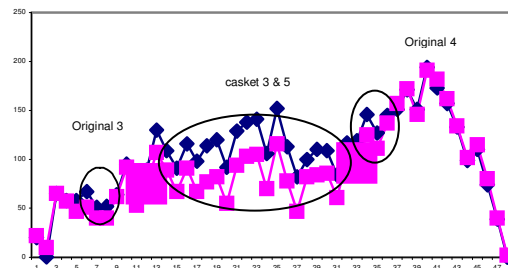


Figure 5a&b: Qsum charts of selected letters as stated.

On the qsum charts of Queen Mary's all original and casket letters both set of charts do not confirm single authorship within them and there is not any way of ruling out any of the possible reasons, whether due to her state of mind or translator issue. The Qsum chart of all casket letters together better illustrated the mixed authorship of the writings, same with Mary's original letters altogether and individually. The most fitting reason for this could be that none of these letters was originally written in English, all of them are translations from Scottish except the original letter 4, which is from French. The qsum charts are reflecting the styles of translators and not the original authorship.

5. Conclusion

In summary, through the blended approach to resolving authorship attribute analysis of stylometric and statistical feature analyses developed in this study, the similarity between the original (1 to 4) and casket letters (one, three and five), and the reasons why it cannot be confirmed was established. All the letters are translations, possibly reflecting the translators' styles and this need to be eliminated before further investigation into resolving the task. The clustering of function word frequency when carried out with insufficient number of words may cluster same genre texts together despite different authorships. To overcome this issue (Binongo 2003) used larger text chunks, which was not possible in this study, all disputed and undisputed letters consist of less than 4000 words. Qsum charts despite their reliability referred to as fingerprints in forensics, were not successful in this study as all of Queen Mary's letters undisputed and disputed are translations and the charts show the translators' style, not the original authorship.

6. Future work

Of the two possible approaches, function word frequency analysis of uniform chunks (200-500 words) of the original and casket letters would invariably reflect the translators' styles, hence not recommended. However, analysis of function words and qsum chart of original letters in Scottish/French could help in resolving the authenticity of Queen Mary's disputed letters as Scottish alphabets are same as that of English.

7. Acknowledgements

Catherine Masi, UC Santa Barbara is thanked for her computer programming contribution.

8. References

- Aaronson, S. (1999). Stylometric Clustering, A comparative analysis of data-driven and syntactic features. www.cs.berkeley.edu/~aaronson/sc/report.doc
- Binongo, J. N. G. (2003). "Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution." *Chance: A Magazine of the American Statistical Association* **16 No.2/Spring 2003**: 9-17. www.stat.duke.edu/chance/162.complete.pdf
- Buckland, W. (1999). "Forensic Semiotics; Review on Analysing for Authorship: A Guide to the Cusum Technique by Jill M. Farrington, with contributions by A.Q. Morton, M.G. Farrington, and M.D. Baker. Cardiff: University of Wales Press, 1996, xii +324pp." *The Semiotic Review of Books* **1999, 10(3)**. www.chass.utoronto.ca/epc/srb/srb/foresem.html
- Farrington, J. M. (2004). "How to be a Literary Detective: Authorship Attribution A brief introduction to cusum analysis." <http://members.aol.com/qsums/QsumIntroduction.html>
- Holmes, D. I. (1998). "The Evolution of Stylometry in Humanities Scholarship." *Literary and Linguistic Computing* **13(3)**: 111-117.
- Lowe, D. and R. Matthews (2005). "Shakespeare vs. Fletcher: A stylometric analysis by radial basis functions " *Computers and the Humanities* **29, Number (6, December, 1995)**: 449-461.
- Merriam, T. (1998). "Heterogeneous Authorship in Early Shakespeare and the Problem of Henry V." *Literary and Linguistic Computing* **13(1)**: 16-28.
- Merriam, T. (2004). "Tamburlaine stalks in Henry VI." *Computers and the Humanities* **30, Number (3, May 1996)**: 267-280.
- Millican, P. (2003). *The Signature Stylometric System: A User-Friendly System for Textual Analysis*. www.etext.leeds.ac.uk/signature/#Documentation
- Smith, P. (2002). *Stylometric Analysis Using Discriminant Analysis: A Study of Sherlock Holmes Stories*. *New Directions in Humanities Computing: Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities, Department for Literary and Documentary Data Processing (LDDV) of the University Computing Center (ZDV) and the Department of English and American Literature, University of Tübingen*. www.uni-tuebingen.de/cgi-bin/abs/abs?propid=27