The Anonymous 1821 Translation of Goethe's Faustus: A Cluster Analytic Approach

Refat Aljumily[1]

[1] Newcastle University

Received: 10 April 2015 Accepted: 4 May 2015 Published: 15 May 2015

Abstract

The scholars, Frederick Burwick and James McKusick, published at Oxford University Press, Faustus from the German of Goethe translated by Samuel Taylor Coleridge in 2007. This edition articulated the result that Samuel Taylor Coleridge is the actual translator of the anonymously published translation Faustus from the German of Goethe (London: Boosey: 1821). The present article tests that result. The approach to test this result is stylometric. Specifically, function word usage is selected as the stylometric criterion, and 80 function words are used to define a 73-dimensional function word frequency profile vector for each text in the corpus of Coleridge's literary works and for a selection of works by a range of contemporary English authors. Each profile vector is a point in 80-dimensional vector space, and 5 different cluster analytic methods are used to determine the distribution of profile vectors in the space. If the result being tested is valid, then the profile for the 1821 translation should be closer in the space to works known to be by Coleridge than to works by the other authors. The cluster analytic results show, however, that this is not the case, and the conclusion is that the Burwick and McKusick result is falsified relative to the stylometric criterion and analytic methodology used. Where, in Popperian terms, falsification does not mean 'prove to be false'. It means that evidence which contradicts a hypothesis has been presented, and it is up to the proposer of the hypothesis either to show that the evidence is inadmissible or irrelevant, or else to emend the hypothesis accordingly. The rest of the article is organized as follows. In section 1 we give the motivation for doing this work. In section 2 we provide a quick introduction to the 1821 Faustus translations that we hope will shed some light on the problem. In section 3 we discuss the previous attempts to attribute the 1821 Faustus to Coleridge. In section 4 we outline the methodology used to add

Index terms— vector space, PCA, MDS, SOM, Isomap, Centroid, variance, stylometric, authorship verification.

The approach to test this result is stylometric. Specifically, function word usage is selected as the stylometric criterion, and 80 function words are used to define a 73-dimensional function word frequency profile vector for each text in the corpus of Coleridge's literary works and for a selection of works by a range of contemporary English authors. Each profile vector is a point in 80-dimensional vector space, and 5 different cluster analytic methods are used to determine the distribution of profile vectors in the space. If the result being tested is valid, then the profile for the 1821 translation should be closer in the space to works known to be by Coleridge than to works by the other authors. The cluster analytic results show, however, that this is not the case, and the conclusion is that the Burwick and McKusick result is falsified relative to the stylometric criterion and analytic methodology used. Where, in Popperian terms, falsification does not mean 'prove to be false'. It means that evidence which contradicts a hypothesis has been presented, and it is up to the proposer of the hypothesis either

to show that the evidence is inadmissible or irrelevant, or else to emend the hypothesis accordingly. The rest of the article is organized as follows. In section 1 we give the motivation for doing this work. In section 2 we provide a quick introduction to the 1821 Faustus translations that we hope will shed some light on the problem. In section 3 we discuss the previous attempts to attribute the 1821 Faustus to Coleridge. In section 4 we outline the methodology used to address the 1821 Faust translation authorship debate. In section 5 we present data preparation. In section 6 we present our main analytical arguments deriving the evidence to refute Coleriadge'a authorship of Faustus. We also present the clustering results obtained in section 6. In section 7 we provide additional interpretation for the analytical results obtained in section 6.We conclude in section 8 with a summary of the results, and discussing open questions and possible future directions.

I. I began to read the book as one who was convinced that the Burwick and McKusick's evidence was sufficient to attribute the translation to Coleridge and, as a stylometrist whose concern is largely methodological, to look closely at the stylometric section (2007: 311-30). I finished it with the conviction, though I am not the first to point it out, that there are grounds for doubt. The analysis was partial and many attribution questions, which I became fascinated with, remained open.

1 Motivation

McKusick's general approach was to use quantitative evidence based on formal indicators of texts, which is in my view, is a correct and instructive methodology. But it was obviously not possible to give a definitive answer to the question of Coleridge's involvement in the translation of Faust. This is the central inquiry of this article.

Given the methods used in his analysis, McKusick drew reasonable conclusions though the methods were insufficient to give more than indicative, that is, inconclusive results. To his credit, McKusick was aware of this and made it clear that the conclusion was suggestive only. McKusick, however, encourages scholars and stylometrists (2007: 315-16, 327, 330) to pursue further analysis and examine the attribution questions raised by the Faust translations, together with the hypothesis advanced in his and Burwick's edition, by using more advanced stylometric methods.

McKusick's approach, however, inspired me to contribute with further evidence to the current literature about the Faust-Coleridge authorship question. In the end my conclusion is quite different. It is based on more advanced multivariate analytical methods, a large number of variables proposed as distinguishing features, and hundred texts. More is said about these in the subsequent discussion.

The scope of my empirical approach is extensive. I have examined not only Coleridge's and other likely candidates' involvement in the translation of

2 Introduction

Goethe published his Faust, the first part of the drama, in 1808. The play attracted considerable publishing interest and publishers of English translations of German's literature decided to translate and publish the play and make extracts from of it available to English readers. Over six partial English translations were issued in about the same time; i.e. the first probably in 1813.

3 III.

Previous Attempts to Attribute the 1821

4 Faust to Coleridge

The 1821 Boosey translation has been variously attributed to the translator of Staël's version (Francis Hodgson), George Soane (1820, 1821, and 1825), John Anster (1820), Daniel Boileau (1820), Leveson Gower (1823), and, recently but strongly, to Samuel Taylor Coleridge (1821). The current scholarly consensus is that none of these translators ever claimed to be the author of Boosey's 1821 edition of Faust.

Paul Zall, a scholar of English Romanticism and American literature, was one of the first researchers to suggest in 1971 a connection between Coleridge and the 1821 translation of Faustus. He observed stylistic similarities between the 1821 Faust and Coleridge's two tragedies, namely Remorse (1813) and Zapolya (1817), and also he sensed echoes of Coleridge's mastery of blank verse in the translation. Literary scholars of the time were not satisfied with the claiming that Coleridge actually translated Faust in 1821. They argued that the case for Coleridge could not be accepted on the available evidence; a great deal of instinct and intuition was used to support the case for Coleridge. To accept it, additional compelling proof should be reached. Following Zall's attempt, Frederick Burwick joined McKusick to re-examine Zall's conclusion with much greater detail. The two scholars make their case that Coleridge was the author and the result included in the 2007 edition referred to above. However, this edition has been much debated and the stylometric analysis has been called into question by many reviewers.

Details of which are available in Goethe's Faust/Coleridge as translator of Goethe's Faust.

In this edition, Burwick's case is based on two types of argument (i) circumstantial historical evidence and (ii) qualitative stylistic criteria, and these are available in **??**1: xv-xxxv). On the other hand, McKusick's case is based on quantitative stylistic criteria, that is, stylometry. The general nature of the article is stylometric and, for this reason, the reminder of the section will focus exclusively on McKusick's stylometric analysis that included in the 2007 edition.

McKusick's role was to find quantitative evidence in support of the joint claim of Coleridgean authorship (1: 312-30). To this end, he compiled a digital electronic corpus comprising: Two types of data were abstracted from the texts comprising the corpus: i) Relative frequencies of word lengths. ii) Relative frequencies of 10 selected function words.

For (i), McKusick counted all two-letter words, all three-letter words, and so on up to eight-letter words for each of the Faust translations and for each of Coleridge's four plays and plotted the word-length frequency distribution for each of these relative to the distribution of the 1821 Faustus. He then applied the chisquared test in order to determine whether or not the differences between the word-length distributions for the anonymous 1821 Faust on the one hand and the five other translations and Coleridge's plays on the other were statistically significant, reasoning that if the differences were significant, then the author of the 1821 Faust could not be the author of the other texts in the corpus. The finding was that the differences between the 1821 translation and Coleridge's Remorse were not significant, but that the differences between the 1821 translation and all the other texts were. His conclusion was that, although such analysis of relative word length frequency "is no longer considered definitive or particularly reliable by stylometrists, it is nevertheless possible to gain interesting and suggestive results by looking at this kind of data" (p.316), and that "although these are not definitive results, they are indeed suggestive. These findings suggest that there is a general similarity in vocabulary, as reflected in wordlength distribution, between Remorse and the 1821 Faustus. There is no such resemblance between the 1821 Faustus and any one of the other contemporary translations of Faust. This finding is consistent with our hypothesis that Coleridge is the author of the 1821 Faustus, and our findings also suggest that, of all of Coleridge's dramatic works, Remorse is the one that most closely resembles the 1821 Faustus in its vocabulary" (p.318).

For (ii), McKusick identified a set of 10 function words, counted their frequencies in each of the texts in his corpus, and then proceeded as for (i) above: the distribution for the 1821 Faustus was graphed and compared to the graphs for each of the other texts, and the differences between each textual pair were tested for statistical significance. And, again as in (i), no significant difference was found between the 1821 Faustus and Remorse, but the differences between Faustus and the other texts were significant.

The conclusion was that "on the basis of the relative frequency of these ten keywords, none of the other contemporary translators is a likely candidate for authorship of the 1821 Faust" (p.327) and that "this finding does not 'prove' that Coleridge is the author of the 1821 Faustus, but this finding is fully consistent with that hypothesis, and (in the absence of other strong contenders) it does indicate a strong likelihood that Coleridge is the author" (p.325).

Speaking about this, McKusick's quantitative stylometric argument supports the case for Coleridge's authorship of the 1821 Faustus, but only weakly. Average word length is an intuitively attractive stylistic criterion, but one whose effectiveness in characterising authorial style and in distinguishing one author from another is at the very least not demonstrated, and there are indications that it is in fact ineffective. McKusick explicitly recognised this in the relevant foregoing quotation, and only went so far as to say that the "general similarity in vocabulary, as reflected in wordlength distribution, between Remorse and the 1821 Faustus" is "suggestive". Function word distribution is a much better stylistic criterion, but Mckusick again claims only that it does not "prove" Coleridge's authorship, but is only "consistent with" it. McKusick appears to realise that the real problem lies not in the selection of stylistic criteria, fundamental as this is, but with logic. A statistically significant difference between two texts relative to some given criterion tells one only that the texts are different, not that they are by different authors, and a statistically non-significant difference that the texts are similar in terms of that criterion, but not that they are by the same author. McKusick's results can only serve to support Coleridge's authorship in this instance. He is thus right in claiming only that his results are "consistent with" the hypothesis of Coleridge an authorship, but his further claim that they indicate a strong likelihood" of it is unjustified.

Overall, therefore, the view of the present article is that McKusick goes beyond the evidence in the title of their re-edition of the 1821 Faustus: From the German of Goethe Translated by Samuel Taylor Coleridge, and this motivates the present discussion to test the result of Coleridge's authorship.

5 IV.

6 Methodology

The present article is concerned specifically with authorship verification (2,3,4): Given a disputed text and a corpus of works by that author, the aim is to decide whether he or she wrote the text. In the present case, this becomes: Is Coleridge the author of the 1821 Boosey translation of Goethe's Faust?

The answer to this question is based on falsifiable methodology. This methodology approaches the problem not by proposing and attempting to justify McKusick's result that Coleridge was or was not the author, but by testing an existing one: the Burwick and McKusick result that he was.

7 b) Principal Components Analysis

PCA is a non-hierarchical linear method based on preservation of data variance. The principal components analysis was in a four-stage procedure. The first step was the construction of a symmetric proximity matrix for distances among vectors. The second was the construction of an orthogonal basis for the covariance matrix in such a way that each axis was the least-squares best fit to one of the n directions of maximum of variation in D. The third was the selection of dimensions; we removed the axes along which that had Volume XV Issue XI Version I Hierarchical cluster analysis constructs clusters in terms of measures of spatial distance among data vectors in the space as the basis for clustering. It provides more information than non-hierarchical ones in that it not only identifies the main clusters, but also its constituency relations relative to one another as well as their internal structures (5,6,7). The hierarchical analysis was in a three-stage procedure. The first step was the calculation of the distances between all possible pairs of vectors. The second was the construction of a onedimensional symmetric matrix of the distances calculated in the first step. The third step was the construction of a hierarchical tree based on the symmetric matrix of distances.

Multivariate methods are used to achieve this. Multivariate methods are essentially variations on a theme: cluster analysis. Cluster analysis aims to detect and graphically to reveal structures or patterns in the distribution of data items, variables or texts, in ndimensional space, where n is the number of variables used to describe an author's style. The class of methods for doing so all depend on finding structure in a highdimensional data space, and then using that structure either to formulate or, in the present case, to attempt to falsify McKusick's result. This class includes hierarchical clustering, principle components analysis, multidimensional scaling, self-organizing map, and Isomap. maximum variation in D, that is, the total combined variance of all vectors (8,9).

8 c) (Metric) Multidimensional scaling

MDS is a dimensionality reduction method which can be used for clustering if the data dimensionality is reduced to three or less. It uses variance preservation as its criterion for keeping as much of the information contained in the original set of data as possible in dimensionality reduction, MDS preserves the proximities among pairs of objects on the basis that the proximity is an indicator of the relative similarities or dissimilarities among the physical objects which the data represents, and therefore of information contained in: if a low-dimensional representation of the proximities can be built, then the representation preserves the information contained in the original data (8,10).

9 d) Self-Organizing Map

SOM has been successfully used in a wide variety of research applications to represent a set of high-dimensional vector points in a low dimensional space without reducing the dimensionality of the original space, while preserving the relationships among the input data vectors. In other words, SOM provides a topology preserving projection from a high-dimensional to a low-dimensional space; that space is usually twodimensional. The property of topology preservation means simply that the projection preserves vector neighborhood relations. Vectors that are near each other in the input space are projected to nearby map units in the SOM. The SOM can therefore be used cluster analysis method by projecting data of arbitrary dimensionality into two-dimensional space and visualizing any structure in the data in a variety of ways (8,11).

10 e) Isomap

Isomap reduces dimensionality by working on a nonlinear rather than on a linear distance matrix. Given a linear distance matrix D L generated from a data matrix M, Isomap approximates the geodesic distances by first deriving a neighbourhood graph to represent different points of a manifold, that is, a geodesic distance matrix D G is approximated mathematically by computing graph distances from D L, and D G is then the ground for dimensionality reduction using either the classical or the metric least squares MDS mathematical procedure. Graph distance approximation to geodesic distance is a widely used paradigm in data analysis to approximate geodesic distance between different points of a manifold using graph distance (8,12).

V.

11 Data Preparation b) Constructing a target corpus and text pre-processing

The standard tradition of creating a corpus for attribution test has always been based on the assumption that the corpus is large and representative of an author respective writings. Therefore, a relevant issue in the current application is what size the corpus should be in order to be representative of Coleridge' literary style. The corpus on which the clustering analysis of Coleridge corpus is based consists of 363 texts of Coleridge's literary output in prose, verse, and drama. However, significant variations in the lengths of these texts are found. Some texts are large enough in size to be analytically practical; they are 31 texts and are shown in Table (1A). Other texts are too short to achieve a good level of analytical accuracy; they are 332 texts and are amalgamated and assigned into 21 collections of texts according to their appearance in journals and poetry collections; they are treated as unitary texts. These are shown in table (1B). In authorship attribution and text clustering, data preparation is the key to obtaining accurate clustering results and to achieve this, variables must be carefully selected. Data analysis should be confined to only and all the important variables that contribute meaningfully to an author's style. In this attempt, the data matrix is built up of only and all the important function words within the texts. The reason for using function words representation is that the frequency distribution of function words is taken to be an indicator of an author's syntactic usage, and, because syntax is largely independent of topic, is regarded as a more reliable criterion for author attribution. Moreover, the experimental results of authorship attribution indicate that function words representation gives good results in identifying the style of a text and distinguishing between a set of authors. Equally important, most studies seem to agree that up till now function word representation has been proven to be giving much better results than any other, more sophisticated stylistic criteria to authorship style. (13,14,15). number of times that function word j occurs in text i. The same is applied to other data matrices (D1, D2, and D3). The texts are given, where necessary, the first or second name or initials as given in the work used. For example Mariner.txt is given for 'the Rime of the Ancient Mariner' and Sibylline.txt is given for 'Sibylline Leaves'. Each matrix row vector therefore represents a function word lexical frequency profile for the corresponding text.

12 Volume XV Issue XI Version I

Since each function word variable in the profile has a label, the profile gives a representation of which function word is in a text and which is not. However, it is observed that the data matrices D, D1, D2, and D3 have some characteristics that can skew the validity of the clustering results. First, there are many super fluousfunction words that are included in the data matrices.

Second, there is a very substantial variation in the lengths of the texts in the data matrices: some texts are very long while others are very short. These matrices have to be transformed prior to analysis.

13 d) Significant and insignificant Function Words

Frequency is the simplest criterion for selecting function words from D, D1, D2, and D3: those function words which occur most often in the texts are judged to be the most important, and those which occur least often are taken to be least important and can therefore be discarded. With respect to clustering, the fundamental idea is that a variable should represent something which occurs often enough for it to make a significant contribution to the clustering of the data vectors. The assumption is that if an individual author uses certain function word frequently in a text, then that function word tells or denotes something about that text or that author's preferred syntactic usage. To select function words based on frequency, given an m x n frequency data matrix D; the value at Dij is the number of times function word j, for j=1?n, occurs in text i, for i=1?m. The frequency of occurrence of function word j across the entire corpus of texts is then:ð ??"ð ??"ð ??"ð ??"ð ??"ð ??"ð ??"ð ??"(????) = ? ?? ??,?? ??=1???

Frequencies for all the columns data matrices (D, D1, D2, D3) are calculated, the function words are sorted in descending order of frequency, the most frequent function words are selected, and the less frequent function words are eliminated from (D, D1, D2, D3). Substantial dimensionality reduction can be achieved by applying this criterion to data matrices (D, D1, D2, D3).

14 e) Text length Normalization

The 52 texts in D, the 53 texts in D1, the 73 texts in D2, and the 23 texts in D3 vary substantially in length. This is shown in Figure (1). The number to the right of each of the text names is the number of words in the text; there is a clear and very strong tendency to cluster by length.

The problem now is that we need a clustering structure that shows the distances among the texts based on the function words similarity, not length. To do this, the row vectors in each data matrix are normalized to adjust the disparity in length among the texts in such a way as to eliminate variation in document length as a factor affecting the frequencies. This normalization is relative to mean document length using the equation: The mean length across all texts are calculated. In each row vector, the count for a function word is multiplied by the mean text length, then divided by the total number of frequency counts occurring in that row vector. The effect of normalization using mean document length is that the values in the row vectors that represent long texts are decreased while the values of the row vectors that represent the short ones are increased. For texts that are near or at the mean, little or no change in the corresponding row vectors occur. The overall effect is that all the corresponding texts are now in effect all the same length and are ready for clustering.

15 f) Data dimensionality and the elimination of low variance variables

Clustering of texts depends on there being variability in their characteristics; identical texts having the same function words cannot be validly clustered.

Where the texts to be clustered are described by function words, then the function words are only useful for the purpose if there is significant variation in the values that they take. In the current application, therefore, we looked for function words with substantial variation in their values, and ignored function words with little or no variation. Function words with no or little variation are removed from data matrices as they contained little information and would complicate cluster analysis by making the data higher-dimensionality than it needs to be. Mathematically, the degree of variation in the values of a variable is described by its variance. The variance of 193 function word values is the average deviation of those values from their mean. The standard definition of variance for an m-row x n-column vectors matrix in which the columns represent 193 function words and the rows represent the texts they describe, the variance of the columns is:

The function word frequencies of the columns in each data matrix are calculated using the above equation and sorted in descending order of frequency magnitude. The column vectors are sorted in descending order as shown in Figure (3). In this figure there are a few relatively highfrequency function words, a moderate number of medium-frequency ones, and a large number of lowfrequency ones. There is considerable scope for dimensionality reduction here; a conservative reduction would be to keep the 80 highest-frequency columns in D, discarding the rest. The same procedure is applied to the other data matrices. That is, function word columns 193 to 80 are removed on the grounds that they contribute little to differentiation of the texts. The selected 80 highest-frequency function words are shown in Table (2): ? = ? = n i i n x v . . 1 2 /)) ((

16 g) Clustering validity

In the present application the generated clustering results are validated in two ways: each data matrix. The clustering analyses of D, D1, and D2 are not shown in this article. There is no hope of being able to show (36) analyses in such an article, but this section addresses them only briefly to the extent to which presentation of the analytical results is necessary for the purpose of this article. The clustering analyses of D showed that there is structure in Coleridge's usage of function words but that usage varies in accordance with genre. The clustering analysis of D1 supports the hypothesis of Coleridge as the author of the 1821 Boosey Faustus, and so is the clustering analysis of D2. This result has serious implications for the validity of the central tent of authorship attribution and the article does not take this similarity as evidence that Coleridge is the actual translator of the 1821 Faustus. This result suggests no more that Coleridge is a likely candidate for the authorship of Faustus since the researcher does not yet know if the five other translations of the play by other likely candidate authors are also closest in style to that of the 1821 text or not. This is where the translations of Faustus by de Staël 1813, Soane, 1821-1825, Anster 1820, Boileau 1820, and Gower 1823 come in. Now all the observations have been captured and the reminder of the discussion will switch to the final stage of the analysis by applying the clustering methods to D3 to see where in the data space the Boosey Faustus sits in relation to the locations of these authors in the space. Because the foregoing clustering results have identified that the Boosey Faustus clusters with closet dramas, and because the additional Faust translations also belong to this genre, only the closet drama (abbreviated CD) texts are clustered and the verse and prose texts are eliminated. This is done for clarity of presentation. D3 contains Faustus, the dramatic texts by Coleridge, Byron, Shelley, Wordsworth, and the translations of Faust by Staël, Soane, Anster, Boileau, and Gower. For this data matrix, we have the following clusterings:

17 T

ii. A range of clustering methods are applied to the same data matrices, each method based on a different view of what constitutes a cluster and how clusters can be identified, and interprets such agreement as is found among them as an indication of the intrinsic or 'true' structure of the data. Specifically:

? PCA is a linear method based on preservation of data variance.

? MDS is a linear method based on preservation of distance relations among objects in data space. ? Isomap is a nonlinear method based on preservation of distance relations among objects in data space.

? SOM is a nonlinear method based on preservation of data topology. ? Single Linkage hierarchical clustering is a linear method based on preservation of data topology. ? Complete, Average, and Increase in Sum of Squares hierarchical clustering are all linear methods based on preservation of distance relations in data space, though they differ in how distance among clusters is defined.

VI.

18 The Clustering Analysis

The data matrices (D, D1, D2, D3) are analysed using five different clustering methods. all of these methods agree with each other in clustering the texts in i. The degree of consistency between the distance matrix underlying the cluster tree and another distance matrix is measured using Cophenetic Correlation Coefficient Measure (5,6,8). Based on this, the trees generated by Average Linkage for D, D1, D2, and D3 seem to fit these data matrices more well than the clusterings produced by Single, Complete, and Ward analyses Coefficient Measure above. In Isomap, CD Faustus is in the neighborhood of Anster, Boileau, and Gower: it is a compromise between Anster Faustus and Boileau's, but far apart from Gower's. Finally, in SOM, CD Faustus is a compromise between CD Anster Faustus and CD Gower Faustus, i.e. it is close to both of them equally.

? Among these authors, the Boosey Faustus is always closer to Anster than to any other author, including

Coleridge. More specifically, Faustus is no longer closest to Coleridge, but to other authors and in particular to Anster and Gower; there's some variation in degree of closeness to these two, but the overall picture is clear.

? No matter how many other authors are included in the test or how many other texts are added to the corpus, that is, more authors or texts won't help: Anster and Gower will always be closer than Coleridge to Faustus.

? Based on the above, therefore, this means that the hypothesis that Coleridge was the author of the 1821 Boosey Faustus is falsified by the methodology used in this test.

Finally, having established that Anster and Gower are closer to Boosey than to Coleridge or any other of the authors included here, it remains to show why, that is, what aspect or aspects of function word usage underlie this result. A centroid-based analysis is used to answer this question. That analysis proceeds as follows.

? From D3, the data matrix used for the preceding cluster analyses, the row representing work by each of the authors are abstracted and, where there is more than one work, the centroid is calculated.

Thus, all the rows of D3 representing work by Coleridge are abstracted and their centroid is calculated, and the same is done for Byron and Shelley; for authors represented by only one work, that is, the various Faust translators and Wordsworth, the corresponding single matrix row is used.

? The set of individual matrix rows and calculated centroids are co-plotted as bar plots and the amount of variation in the variable centroids are calculated. A variable with a larger amount of variability in its centroid than the other variables in a set of data is taken to be the most important discriminator between the authors or the clusters of interest because there is much change in the values of that variable throughout text row vectors.

? Because it is difficult to interpret the very crowded bar plots for the full 80 variables, only the dozen variables with the largest variation in relative bar plot heights are shown in what follows.

The centroids of most important function words to each of the authors are first calculated, as shown in Table **??**3) and the resulting centroids are then bar plotted onto a bar chart, as shown in figure (9): The number and type of function words per column has been represented along the horizontal axis, and the centroids per column up the vertical axis. Each one of the function words has its own a label on the horizontal x-axis that holds a value on the vertical y-axis of the bar chart, where the height of each bar represents the variable centroid containing the values of a given variable in each text row vector. The bars are displayed arbitrarily following the order of the function words, which are given in table (3) rather than ordered by size from the smallest to largest or vice versa.

From Table **??**3) and the plot in Figure (9), it can be seen that there is pattern of differences among the 10 authors considered in the study with respect to the most important functions words and this yields empirically stylistic criteria showing how each author's usage of a set of 10 function words, and, more particularly, how the usage of this set of 10 function words by Anster, Coleridge, the 1821 anonymous translator, and Gower does not overlap with that of each other's or any other author's usage. For example, Staël shows a higher usage of 'of' and 'to' than in any other author, the 1821 anonymous translator shows a higher usage of 'and' than in any other author, Shelley shows a lower usage of 'then' than in any other author, Wordsworth and Boileau show a lower, though an equal, usage of 'yet'. Boileau and Staël show a lower usage of 'or' than in any other author. For others, the usage of this set of 10 function words is somewhere between these extremes. For example, 'of', 'and', and 'to' usages are very frequent in Anster's Faustus; 'of', 'and', 'that', and 'with' usages are much lower in Byron's than in any other author; 'and', 'of', 'to', and 'that' usages are more frequently in Boileau's than in some other authors; 'of', 'and', 'to', and 'that' usages are frequent and consistent in Coleridge's dramas and so are in Wordsworth's The Borderers. The usage of 'then' is much higher in Faustus than in any other author. Finally, 'from', 'or', 'with', and 'by' are marked with relatively consistent or frequent usages among all the authors and therefore do not distinguish between them.

All in all, based on the centroid values in the Table **??**3) above and their corresponding plots in the Figure (9), we can draw the following results:

? Function words 'that', 'and', and 'with' are the most important in determining the distance relations in the foregoing cluster analyses. This is based on the amount of variation in each variable-centroid, which is calculated and shown in varies from the other authors, and in particular from the 1821 anonymous translator, Anster, and Gower in terms of his usage of 'that', 'to, 'then', 'from', 'and', and 'of', which is either higher or less than them. This is a substantive, empirically-based criterion for distinguishing the styles of the authors which have been included in the study, with respect to the closet drama genre. The general conclusion is that the 1821 Faust translation is mathematically similar to the translations of the play by Anster and Gower and that the function words 'of', 'yet' and 'that' are the main determinants for that similarity. This is a plausible result for Anster and Gower, but it is by far not the only interpretation. The next section will justify this claim.

19 VII. Additional Interpretation

Since all of the three translations appear in such close proximity, the conclusion would surely be that either Anster or Gower translated the 1821 Faustus (Boosey edition); or at least that Anster and Gower are likely the best candidates for its authorship, considering Anster as the most probable translator among the translators tested and Gower among the less likely. In such a case, the question is: can the 1821 anonymous Faustus be attributed to Anster or should it rather be attributed to Gower based on this new evidence? The answer is no. The argument is that it is perhaps not so surprising that the 1821 Faustus, claimed by Burwick and McKusick for Coleridge, is closer to two other contemporary translations of the play by Anster and Gower. There are only a limited number of function words that can be used to translate the German words of the original; and the possibility of borrowing from one author to another is also stronger. Many examples could be given of such borrowing of function words (and other style features), but few will suffice here to support this claim. These are taken from Anonymous (trans.) Faustus from the German of Goethe. London: Boosey and Sons, 1821; John Anster (trans.) 'The Faustus of Goethe', Blackwood's Edinburgh Magazine, vii, 1820; and Leveson-Gower (trans.) Faust: A Drama By Goethe. They are quoted, identified by the verse lines, and then highlighted.

20 Line number

Anster Anster and Gower: specific function words and (short phrases) used by Anster were used by the anonymous translator of the 1821 Faustus and Gower as well as some function words used by the anonymous translator of the 1821 Faustus were used by Gower in his own translation (though Gower borrowed less frequently than the 1821 anonymous translator). And this has the effect of clustering the three translations by Anster, the anonymous translator, and Gower together.

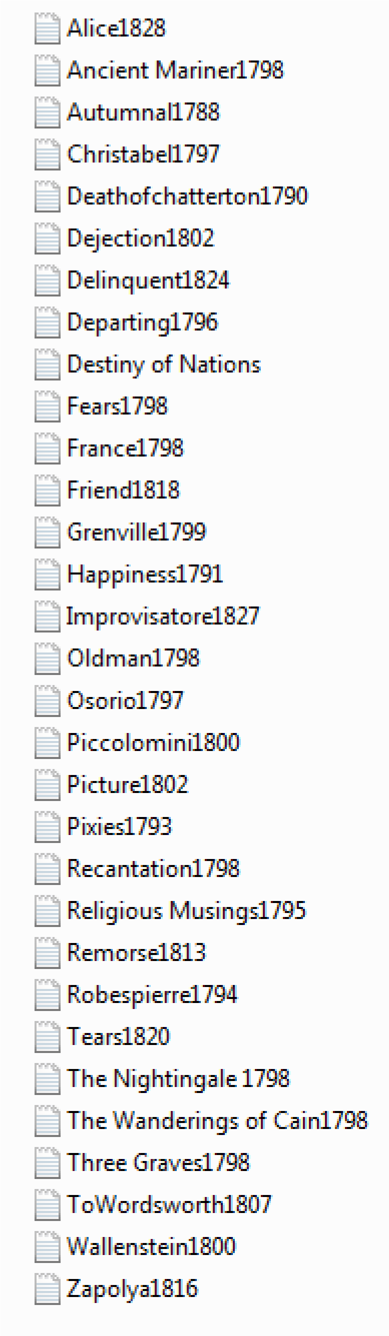
The historical and, to some degree, the literarycritical evidence suggest Coleridge an authorship, but the stylometric evidence, based on what is currently regarded as the best stylometric criterion and using objective and replicable mathematical methods, suggests otherwise. The study has analysed Coleridge's plays and has found they are mathematically quite distinct from the 1821 Faustus translation. However, it is important not to over-interpret this result since the present attribution attempt is based on a particular type of test, proximity in vector space, using a particular stylistic criterion, the frequency of function word usage. Other stylistic criteria and/or other types of test may well give a different result, and the next research step with respect to the Burwick and McKusick result is to devise other types of test based on other criteria. Any future study must, however, take account of the result of the present one, and until one or more such studies appear, the Burwick and McKusick result is abandoned. The article also has closely examined the Faust text and the texts by the 1821 anonymous translator of the 1821 Faust, Anster, and Gower and found that translating the words of the original text of Faust slides over into borrowing from one author into another. [1] [2]

2007

Figure 1: n 2007 ,

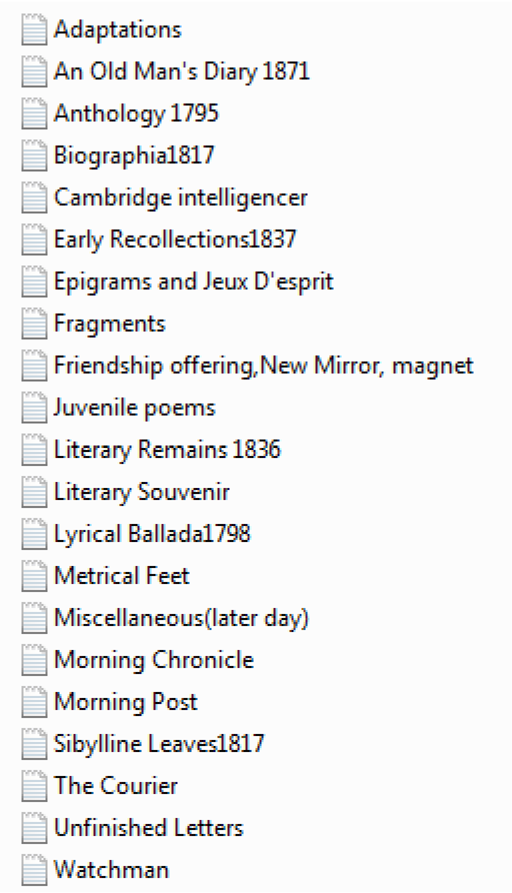
[1] © 2015 Global Journals Inc. (US)

[2] © 2015 Global Journals Inc. (US)



Alice1828
Ancient Mariner1798
Autumnal1788
Christabel1797
Deathofchatterton1790
Dejection1802
Delinquent1824
Departing1796
Destiny of Nations
Fears1798
France1798
Friend1818
Grenville1799
Happiness1791
Improvisatore1827
Oldman1798
Osorio1797
Piccolomini1800
Picture1802
Pixies1793
Recantation1798
Religious Musings1795
Remorse1813
Robespierre1794
Tears1820
The Nightingale 1798
The Wanderings of Cain1798
Three Graves1798
ToWordsworth1807
Wallenstein1800
Zapolya1816

Figure 2:



- Adaptations
- An Old Man's Diary 1871
- Anthology 1795
- Biographia1817
- Cambridge intelligencer
- Early Recollections1837
- Epigrams and Jeux D'esprit
- Fragments
- Friendship offering,New Mirror, magnet
- Juvenile poems
- Literary Remains 1836
- Literary Souvenir
- Lyrical Ballada1798
- Metrical Feet
- Miscellaneous(later day)
- Morning Chronicle
- Morning Post
- Sibylline Leaves1817
- The Courier
- Unfinished Letters
- Watchman

Figure 3:

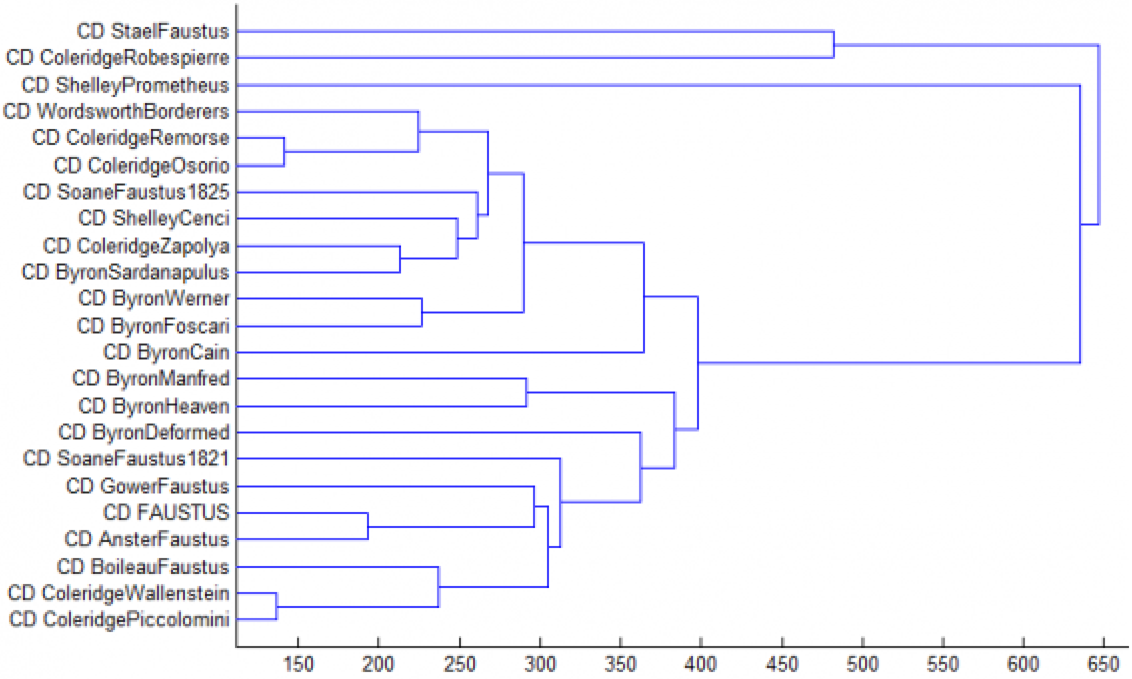
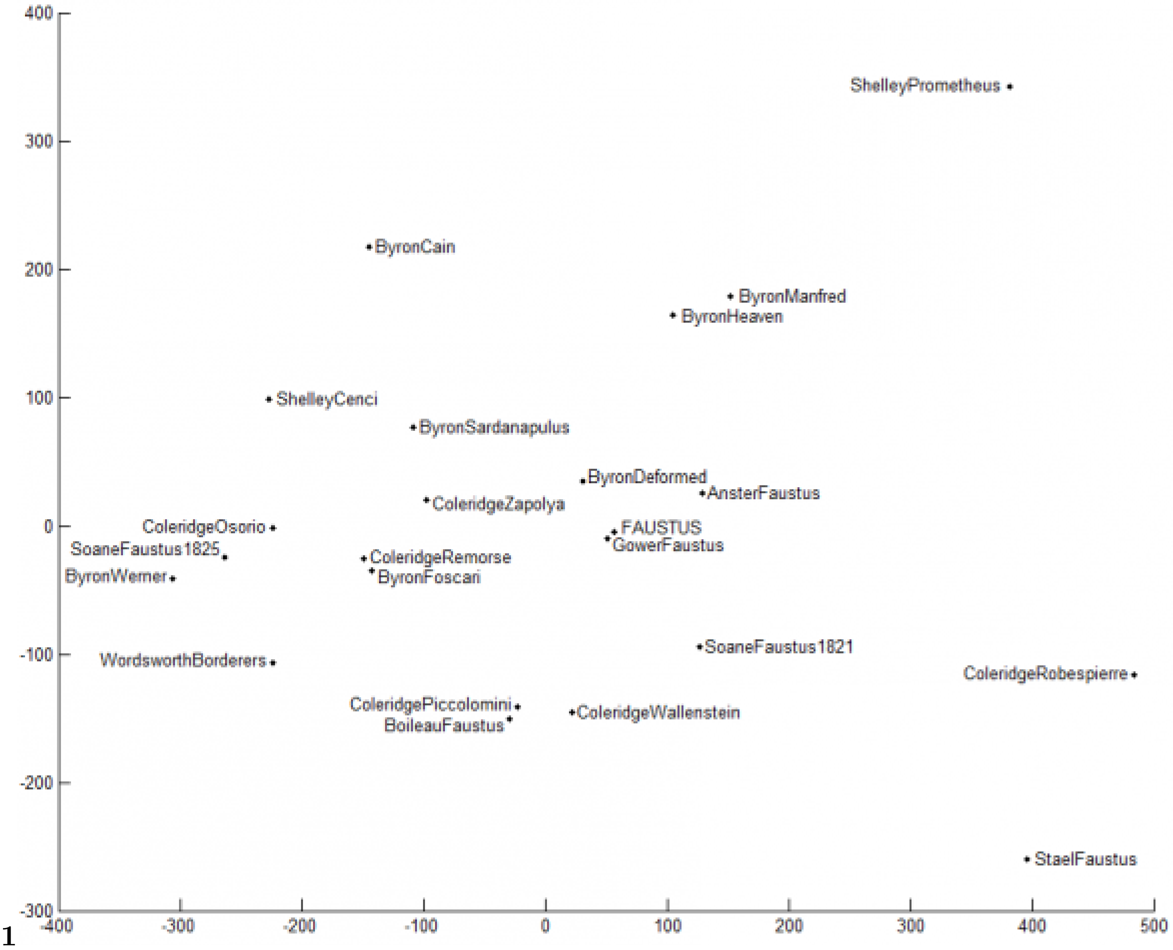
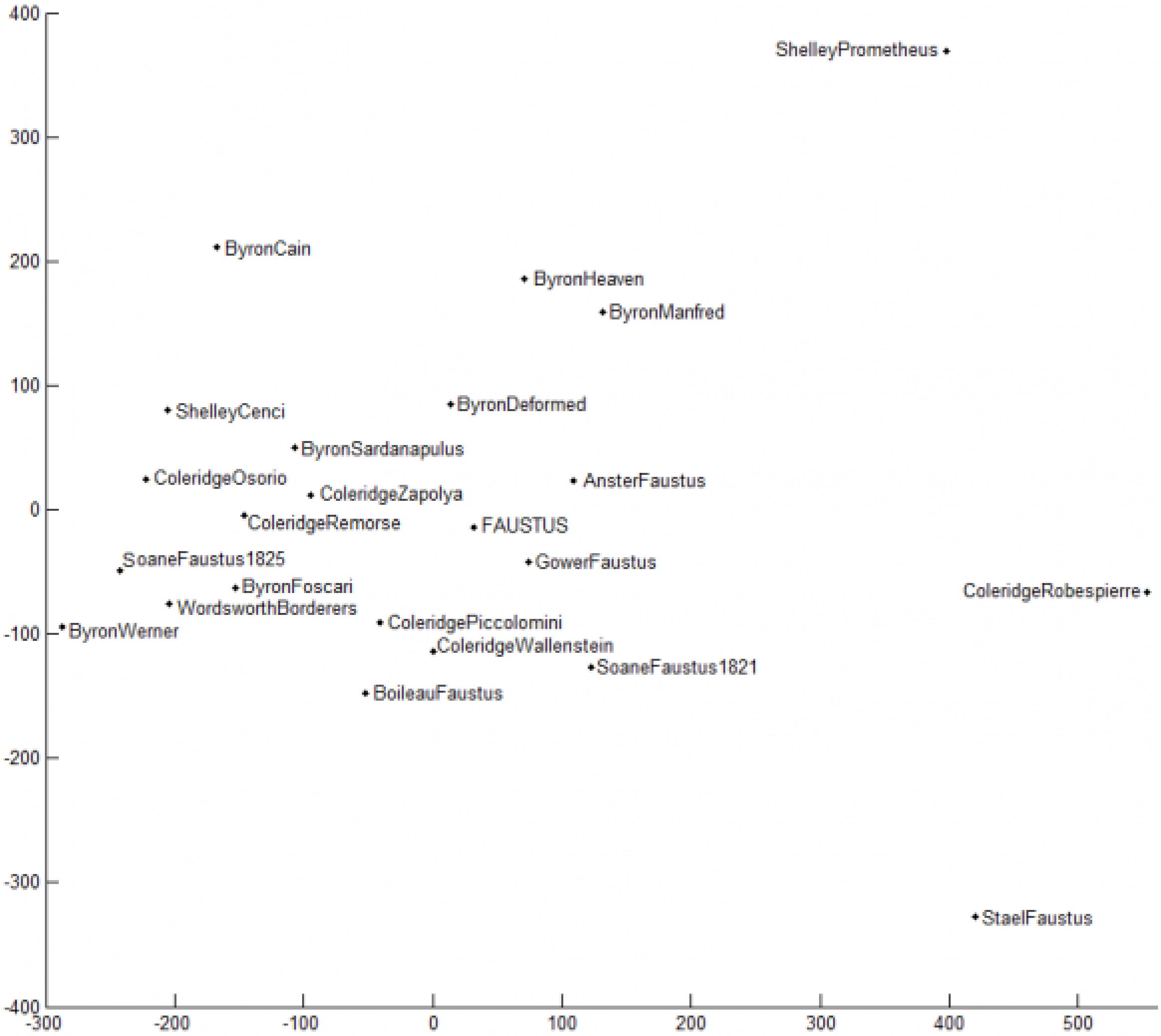


Figure 4:



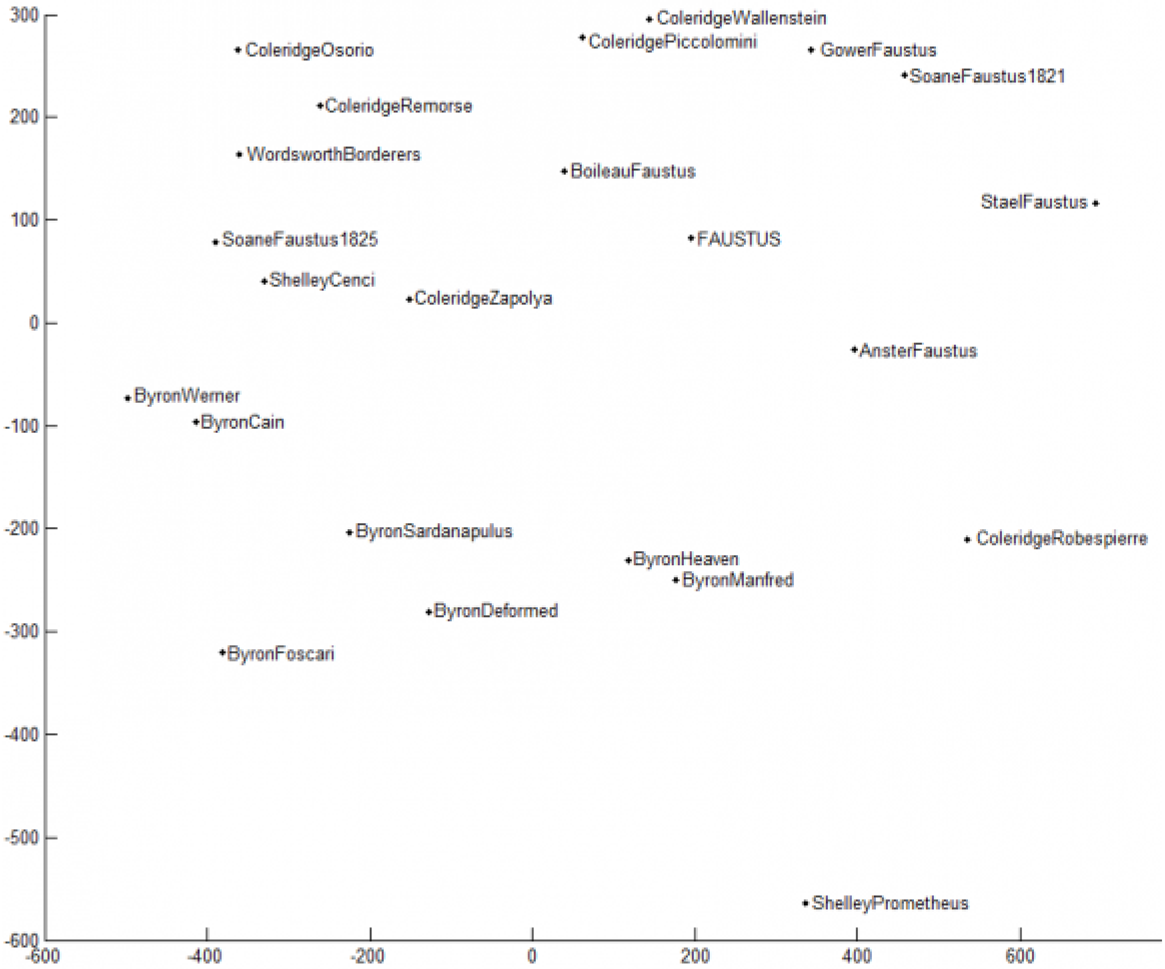
1

Figure 5: able 1 :



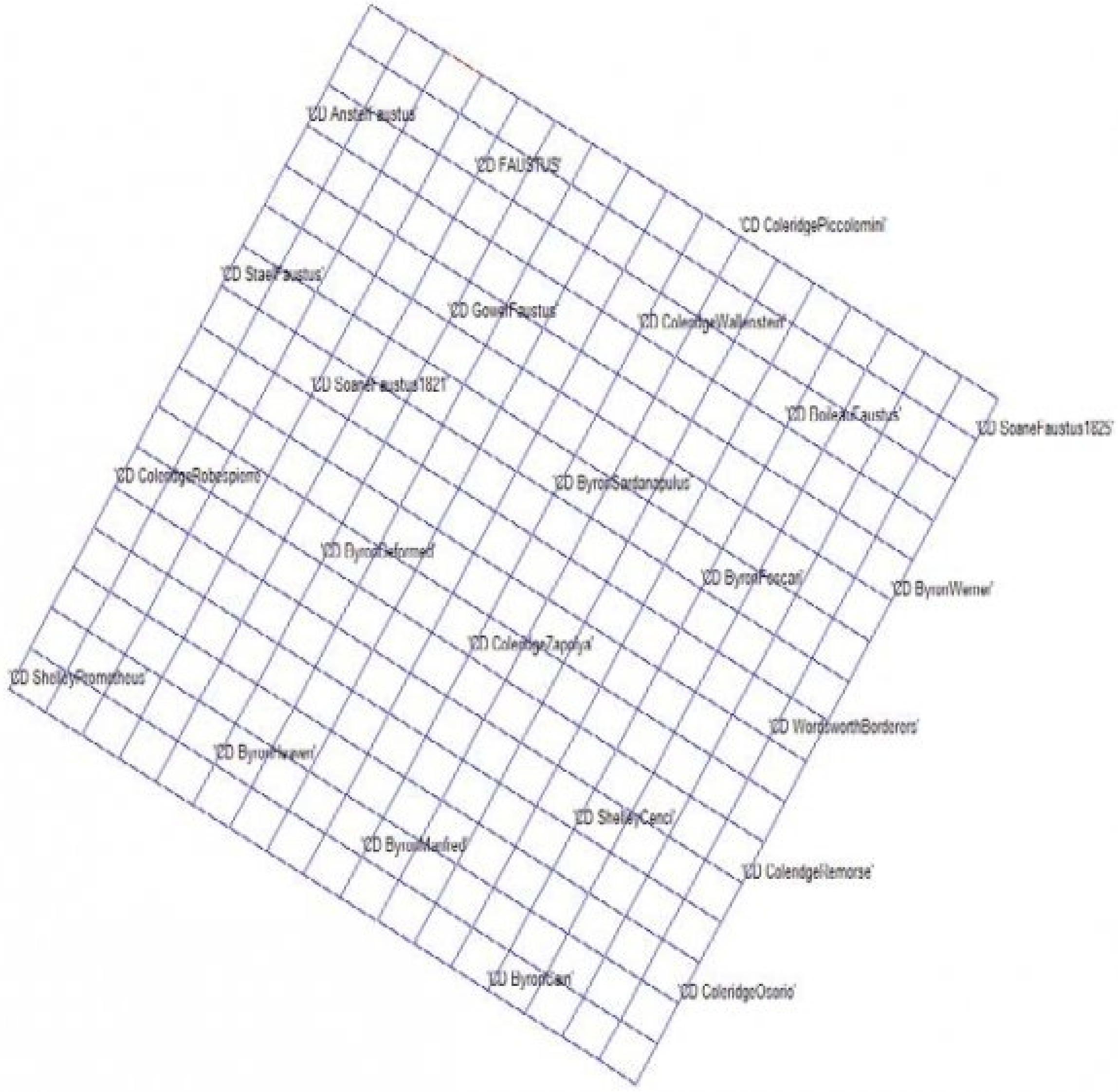
1

Figure 6: Figure 1 :



ColeridgeWallenstein
ColeridgePiccolomini
ColeridgeOsorio
GowerFaustus
SoaneFaustus1821
ColeridgeRemorse
WordsworthBorderers
BoileauFaustus
StaelFaustus
SoaneFaustus1825
FAUSTUS
ShelleyCenci
ColeridgeZapolya
AnsterFaustus
ByronWerner
ByronCain
ByronSardanapulus
ColeridgeRobespierre
ByronHeaven
ByronManfred
ByronDeformed
ByronFoscari
ShelleyPrometheus
300
200
100
0
-100
-200
-300
-400
-500
-600
-600
-400
-200
0
200
400
600

Figure 7:



2

Figure 8: Figure 2 :

Figure 9:

3

Word type	Anster	Boileau	Byron	Coleridge	Faustus	Gower	Shelley	Stael	Soane	W.worth
of	475	363	213	381	400	293	315	733	316	338
from	115	75	45	88	103	85	64	81	90	76
or	49	26	43	36	33	56	45	28	46	39
and	585	508	308	477	601	533	407	413	470	447
with	176	156	75	150	169	154	90	147	158	104
then	35	35	21	48	71	40	12	26	83	29
yet	30	21	25	44	33	74	23	22	45	21
To	406	433	208	357	428	445	168	560	365	381
by	80	57	34	62	55	58	39	78	79	69
that	181	152	84	192	167	133	105	220	165	226

Figure 10: Table 3 :

4

? Function words 'and' and 'with' are those with respect to which Anster and the 1821 anonymous translator are closest, and 'with' is that to which Gower and the 1821 anonymous translator are closest.
? Coleridge's usage of this set of 10 function words

Figure 11: Table 4 :

Word type	Amount of variation
of	19.9977.1222
from	379.7333
or	90.3222
and	7733.2111
with	1226.5444
then	487.3333
yet	280.1777
to	13050
by	256.9888
that	2114.0555

	354-364 1675-1682	1820 Alas! I have explored Philosophy, and law, and medicine, And over deep divinity have pored, Studying with ardent and laborious zeal Andhere I am at last, a very foal, With useless learning cursed, No wiser than at first! They call me doctor-and I lead These ten years past my pupils' creed, What can'st thou give, poor miserable devil.	Anonymous 1821 Now I have toil'd thro' all; philosophy, Law, physic, and theology: alas All, all I have explor'd; and here I am A weak blind fool at last: in wisdom risen No higher than before: Master and Doctor They style me now; and I for ten long years Have led my pupils up and down, thro' paths Involv'd and intricate, only to find Thou miserable fiend? can man's high spirit,	Gower 1823 WITH medicine and philosophy I have no more to do; And all thy maze, theology, At length have waded through And stand a scientific fool, As wise as when 1 went to school. 'Tis true, with years of science ten, A teacher of my fellow men, Above, below, and round about, Not Translated
		Thinkest thou that man's ?By suchasthou art? wretch, what canst thou give?	Full of immortal longings, be by such As thou art, comprehended? Thou	

.1 Acknowledgments

The author wishes to thank all those who dedicated their time answering my queries and providing me with valuable comments during the preparation of this study.

.2 a) Conflicts of Interest

The author declares no conflict of interest.

[Everitt et al. ()] , B S Everitt , S Landau , M Leese . 2001. Arnold: London.

[Argamon and Levitan (2005)] , S Argamon , S Levitan . http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.71.6935&rep=rep1&type=pdf 2005. July 2009. p. 11.

[Hair et al. ()] , J Hair , W Black , B Babin , Anderson , R Multivariate Data . *Analysis* 2010. Prentice-Hall International. (7) . (th ed)

[Koppel et al. (2008)] , M Koppel , J Schler , Online . http://www.machinelearning.org/proceedings/icml2004/papers/415.pdf July 2008. p. 20.

[Anderberg ()] *Cluster analysis for applications*, Michael Anderberg . 1973. Academic Press, Inc: London.

[Moisl ()] *Cluster analysis for corpus linguistics*, Hermann Moisl . 2015. Berlin: De Gruyter Mouton.

[Romesburg ()] *Cluster Analysis for Researchers*, Charles Romesburg . 1984. Wadsworth Inc: USA.

[Koppel and Schler ()] 'Computational methods in authorship attribution'. Moshe Koppel , J Schler , Argamon , S . *Journal of American Society for Information and Technology* 2009. 60 p. .

[Burwick and Mckusick ()] *Faustus from the German of Goethe Translated by Samuel Taylor Coleridge*, Frederick Burwick , James Mckusick . 2007. Great Britain: Oxford University Press.

[Juola] Patric Juola . *Authorship attribution. ACMDL2006*, 1 p. .

[Bishop ()] *Neural Networks for Pattern Recognition*, C M Bishop . 1995. USA: Oxford University Press.

[Lee and Verleysen ()] 'Nonlinear Dimensionality Reduction'. J A Lee , M Verleysen . *Springer science and business media*, (New York) 2007.

[Grieve ()] *Quantitative authorship attribution: an evaluation of techniques. Literary and linguistic Computing*, Jack Grieve . 2007. 22 p. .

[Kohonen ()] *Self-Organizing Maps*, Teuvo Kohonen . 2001. Berlin: Springer. (3rd ed)

[Holmes ()] *The Evolution of Stylometry in humanities scholarship. Literary and Linguistic Computing*, David Holmes . 1998. 13 p. .