

**The anonymous 1821 translation of Goethe's *Faustus*: A
cluster analytic approach**

**By
Refat A. Ali**

**A thesis submitted in fulfilment of the requirements for
the degree of Doctor of Philosophy
School of English Literature, Language and Linguistics
Newcastle University**

September, 2015

Abstract

This study tests the hypothesis proposed by Frederick Burwick and James McKusick in 2007 that Samuel Taylor Coleridge was the author of the anonymous translation of Goethe's *Faust* published by Thomas Boosey in 1821. The approach to hypothesis testing 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 cluster analytic methods are used to determine the distribution of profile vectors in the space. If the hypothesis 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 hypothesis is falsified relative to the stylometric criterion and analytic methodology used.

Dedication

*To my aged mother,
Without your blessings on me none of my success would be
possible*

Acknowledgments

Thanking every individual without missing others is a daunting task. But, I would like to express my sincere gratitude and thank all individuals who have supported me during these years and who have become invaluable for me along this Ph.D. journey, to only some of whom it is possible to give a particular mention here.

Above all, this thesis would be unthinkable without the commitment of Dr. Hermann Moisl, my retired teacher and principal supervisor. He pushed me to a higher level of stylometric research, by emphasizing the importance of quantitative methodology and innovation, but also by having confidence in me. He was always there when needed, mathematical training and computational techniques, always ready with solutions for the problems, and always tolerant through the downs in the research, which God Almighty knows how frequent they were, and I cannot thank him enough for all that he has done for me.

I will forever be thankful to my second teacher and supervisor Prof. Michael Rossington, firstly, for suggesting the topic of this thesis and secondly, for his insightful comments for improving the materials related to the literary side of the thesis and for tracking down some elusive information which led to the development of the ideas presented in it.

My deepest heartfelt appreciation goes to Prof. Charles Romesburg from the Utah State University and Prof. James C. McKusick from the University of Montana for taking time out from their busy schedule to answer my inquires and provide me with valuable remarks.

I am also greatly indebted to all my APR panelists who provided me with useful suggestions and guidelines throughout my Ph.D research project, particularly: a retired teacher Prof. Noel Burton Roberts and Dr. Geoffrey Poole.

I owe a very important debt to Prof. Anders Holmberg who was truly an influential part of my whole Ph.D application process: I would not be here without his indispensable comments on my early preliminary proposal of the thesis and application which led to the interviews then Ph.D course admission.

Thanks are due to the Librarians of Robinson library-Newcastle University, Middlesbrough central library, Grimsby town library, and Immingham library for their help and assistance in providing me with the valuable references and sources needed for the research. Additionally, I would like to thank the directors of postgraduate studies, the former, Dr. James Procter and, the current, Dr. Anne Whitehead, for their advice and assistance.

Last, but by no means least, I would like to thank three important groups of people for their support, encouragement, and love. First and foremost, special thanks to my mother whose prayer requests contributed a lot to my entire life and to the completion of this project in particular. I owe a debt of gratitude to my brother Mr. Wajdi, my best friends Mr. Nick Plummer and Pauline McLaughlin. I am lucky to have met Shelley Gibson here, and I thank her for her love, support, and unyielding encouragement.

Table of Contents

Abstract.....	i
Dedication.....	ii
Acknowledgments.....	iii-iv
Table of Contents.....	v-vii
List of Figures.....	viii-xi
List of Tables.....	xii-xiii
List of Appendices.....	xiv
List of Abbreviations.....	xv
Introduction.....	1-2
Chapter One: Motivation, History and Current State of the 1821 <i>Faustus</i> Translation Authorship Debate	
1.1 Motivation.....	3-4
1.2 Bibliographic overview of translations of <i>Faustus</i> (Part I) in the early nineteenth-century.....	4-6
1.3 Existing attributions of Boosey's 1821 <i>Faustus</i> to Coleridge.....	6-7
1.3.1 The circumstantial historical argument.....	8-11
1.3.2 The qualitative stylistic argument.....	11-15
1.3.3 The quantitative stylistic argument.....	15-18
1.4 Assessment.....	18
1.4.1 The present discussion's own reaction.....	18-24
1.4.2 Other reactions.....	24
1.4.2.1 The circumstantial historical argument.....	24-26
1.4.2.2 The qualitative stylistic argument.....	26-28
1.4.2.3 The quantitative stylistic argument.....	28-31

Chapter Two: Research Question and Methodology

2.1 Research question.....	32
2.2 Methodology.....	32
2.2.1 The authorship identification problem.....	32-33
2.2.2 Literature review.....	33-34
2.2.2.1 Older works.....	34-36
2.2.2.2 Recent developments.....	36-55
2.2.3 The methodology used in the present study.....	56
2.2.3.1 Hypothesis testing.....	56-61
2.2.3.2 Vector space methods.....	61-72
2.2.3.3 Data creation.....	72-99
2.2.3.4 Data analysis.....	99-129

Chapter Three: Analysis

3.1 Data creation: Function words frequency in Coleridge's works.....	131-145
3.2 Coleridge's usage of function words.....	145-156
3.3 Comparison of Coleridge's usage of function words with contemporary authors.....	156-169
3.4 Where <i>Faustus</i> fits.....	169-180
3.5 Coleridge and the other translators of <i>Faustus</i>	180-198

Chapter Four: Interpretation.....199-210

Chapter Five: Conclusions, Limitations, and Further Research

4.1 Conclusions.....	211-215
----------------------	---------

4.2 Limitations.....	215
4.3 Further research.....	215-216
Appendices.....	217-260
Bibliography.....	261-304

List of Figures

Figure (1.1) Word length measurements for <i>Faustus</i> 1821 and Remorse.....	16
Figure (1.2) Word length measurements for <i>Faustus</i> 1821 and Anster's translation.....	16
Figure (1.3) Word length measurements for <i>Faustus</i> 1821 and Gower's translation.....	17
Figure (2.1) An example of a vector	61
Figure (2.2) Data items and variables in a data matrix $m \times n$	62
Figure (2.3) 2- and 3-dimensional vector space.....	64
Figure (2.4) A vector in space.....	64
Figure (2.5) Vector length.....	65
Figure (2.6) The angle between vectors.....	66
Figure (2.7) Vector distances.....	66
Figure (2.8) Figure.....	67
Figure (2.9) Figure.....	67
Figure (2.10) Figure.....	68
Figure (2.11) Euclidean distance measure.....	69
Figure (2.12) Euclidean distance between V_1 and V_2	70
Figure (2.13) Text-length based clustering.....	78
Figure (2.14) Categories of manifold definition.....	83
Figure (2.15) Effect of dimensionality increase on the size of a cube.....	84
Figure (2.16) Data set of 1000 vectors in 3-dimensional space.....	86
Figure (2.17) Plots of very large vectors in 2-dimensional space.....	86
Figure (2.18) Five 2-dimensional vectors in space	88
Figure (2.19) Two 3-dimensnional vectors in space.....	88
Figure (2.20) Sparse data in the space.....	90
Figure (2.21) Concentrations of distance among vectors in space.....	91
Figure (2.22) Scatter plots of 2-dimensional data.....	100

Figure (2.23) Two-dimensional data distribution with orthogonal basis.....	104
Figure (2.24) Alternative orthogonal basis	105
Figure (2.25) Highly correlated two-dimensionality vectors with orthogonal basis.....	105
Figure (2.26) Alternative orthogonal basis for vectors.....	106
Figure (2.27) Three-dimensional data distribution with orthogonal basis.....	106
Figure (2.28) $N \times N$ covariance matrix of 6 phonetic segments for DMC.....	107
Figure (2.29) Linear and non-linear distance between points on the Earth's surface.....	115
Figure (2.30) A manifold embedded in metric space.....	116
Figure (2.31) Neighborhoods in metric space.....	117
Figure (2.32) Scatter plot of randomly generated two-dimensional matrix M	118
Figure (2.33) Graph interpretation of the neighborhood matrix.....	120
Figure (2.34) Structure of a self-organizing map.....	122
Figure (2.35) SOM input lattice.....	123
Figure (2.36) SOM lattice.....	123
Figure (2.37) An example of SOM trained on 20 vectors.....	124
Figure (2.38) Hierarchical clustering tree.....	126
Figure (2.39) Single linkage clustering.....	127
Figure (2.40) Complete linkage clustering.....	127
Figure (2.41) Average linkage clustering.....	128
Figure (3.1) Variation in the lengths of the texts in the Coleridge's Matrix D	142
Figure (3.2) Ward's analysis of Coleridge's Matrix D	143
Figure (3.3) The distribution of function words in frequency matrix $F1$	144
Figure (3.4) Single linkage with cophenetic correlation.....	147
Figure (3.5) Complete linkage with cophenetic correlation.....	148
Figure (3.6) Average linkage with cophenetic correlation.....	149
Figure (3.7) Ward linkage with cophenetic correlation.....	150

Figure (3.8) PCA of M80Norm.....	152
Figure (3.9) MDS of M80Norm.....	153
Figure (3.10) Isomap of M80Norm.....	154
Figure (3.11) SOM of M80Norm.....	155
Figure (3.12) The distribution of function words in frequency matrix F2.....	157
Figure (3.13) Single linkage with cophenetic correlation.....	158
Figure (3.14) Complete linkage with cophenetic correlation.....	159
Figure (3.15) Average linkage with cophenetic correlation.....	160
Figure (3.16) Ward linkage with cophenetic correlation.....	161
Figure (3.17) Unlabeled clustering results.....	163
Figure (3.18) PCA of M180Norm.....	164
Figure (3.19) MDS of M180Norm.....	165
Figure (3.20) Isomap of M180Norm.....	166
Figure (3.21) SOM of M180Norm.....	167
Figure (3.22) Single linkage with cophenetic correlation.....	170
Figure (3.23) Complete linkage with cophenetic correlation.....	171
Figure (3.24) Average linkage with cophenetic correlation.....	172
Figure (3.25) Ward linkage with cophenetic correlation.....	173
Figure (3.26) PCA of M280Norm.....	175
Figure (3.27) MDS of M280Norm.....	176
Figure (3.28) Isomap of M280Norm.....	177
Figure (3.29) SOM of M280Norm.....	178
Figure (3.30) The distribution of function word frequency matrix F4.....	181
Figure (3.31) Single linkage with cophenetic correlations.....	182
Figure (3.32) Complete linkage with cophenetic correlations.....	182
Figure (3.33) Average linkage with cophenetic correlations.....	183

Figure (3.34) Ward linkage with cophenetic correlations.....	183
Figure (3.35) PCA of M380Norm.....	185
Figure (3.36) MDS of M380Norm.....	186
Figure (3.37) Isomap of M380Norm.....	187
Figure (3.38) SOM of M380Norm.....	188
Figure (3.39) Bar plot for 10 authors based on centroid analysis of 10 function words.....	191
Figure (3.40) The usage of the 10 most important function words across <i>Faustus</i> , <i>Piccolomini</i> , and <i>Wallenstein</i>	195
Figure (3.41) The usage of the 10 most important function words across <i>Faustus</i> , Anster, and Gower.....	197

List of Tables

Table (2.1) An example of Type/Token ratio.....	37
Table (2.2) Matrix M of 3 length-varying documents described by 3 variables.....	80
Table (2.3) Matrix M of 3 documents length-normalized frequency profile.....	81
Table (2.4) An example of a mean.....	95
Table (2.5) An example of 10 values for two variables x and y.....	96
Table (2.6) A matrix underlying figure (2.32).....	119
Table (2.7) Shortest-path graph distance table.....	121
Table (3.1) A selection of Coleridge's works.....	131-139
Table (3.2) Small side-by-side sample of original and corresponding cleaned text.....	141
Table (3.3) A list of 193 function words.....	141-142
Table (3.4) A fragment of a 52 x 193 data matrix D.....	142
Table (3.5) Full names and corresponding abbreviations of Coleridge's texts.....	145-146
Table (3.6) Cophenetic correlation for M80Norm.....	151
Table (3.7) A selection of works from Byron, Shelley, and Wordsworth.....	156-157
Table (3.8) Cophenetic correlation for figures (3-13,14,15, 16).....	162
Table (3.9) Cophenetic correlation for M280Norm.....	174
Table (3.10) Types of 80 high-variance words in figure (3.30).....	181
Table (3.11) Cophenetic correlation coefficients for M380Norm.....	184
Table (3.12) Function word frequency centroids for 10 authors	191
Table (3.13) The amount of variation in the centroids of the 10 function words for 10 authors.....	193
Table (3.14) Function word frequency centroid for sub-cluster texts of interest (1) based on 10 FWs.....	194
Table (3.15) The amount of variation in the centroids of 10 function words for <i>Faustus</i> , <i>Piccolomini</i> , and <i>Wallenstein</i>	196

Table (3.16) Function word frequency centroid for sub-cluster texts of interest (2) based on 10 FWs.....	196
Table (3.17) The amount of variation in the centroids of 10 function words for <i>Faustus</i> , Anster, and Gower.....	197-198

List of Appendices

The ASCII texts used in the study

Appendix 1: 363 texts by Coleridge	217
Appendix 2: The 31 long text by Coleridge.....	218
Appendix 3: The 332 short texts by Coleridge aggregated into 21 texts.....	219
Appendix 4: The aggregated 21 texts by Coleridge.....	220
Appendix 5: 10 texts by Byron.....	220
Appendix 6: 6 texts by Shelley.....	220
Appendix 7: 5 texts by Wordsworth.....	221
Appendix 8: 5 texts by other Faust translators.....	221

The programmes used in the study:

Appendix 9: MAT LAB version R2013a.....	221-252
Appendix10: Cluster Analysis version ClustanGraphics3.....	252-254
Appendix11: Clean texts software.....	254-255
Appendix12: Generate frequency matrix.....	255-257
Appendix13: Edit matrix.....	257-259
Appendix14: Centroid vectors.....	259-260

List of Abbreviations:

AA	Authorship Attribution
CA	Cluster Analysis
CW	Content Words
DBSCAN	Density Based Spatial Clustering with Noise
EDA	Exploratory Data Analysis
EMVA	Exploratory Multivariate Analysis
FW	Function Words
IR	Information Retrieval
MVA	Multivariate Analysis
NLP	Natural Language Processing
PCA	Principal Component Analysis
TF-IDF	Term frequency- Inverse Document Frequency
SEE	Sum of Squares Error
SOM	Self Organizing Map
SVD	Singular Value Decomposition
SVM	Support Vector Machines
UPGMA	Unweighted Pair Group Using Mathematic Averages
VAT	Visual Assessment of clustering Tendency
VSM	Vector Space Model
WPGMA	Weighted Pair Group Using Mathematic Averages
WPGMC	Weighted Pair Group Method of Clustering
WSJ	<i>Wall Street Journal</i>

Introduction

Part I of *Faust*, which appeared in 1808, is one of the most celebrated works of Johann Wolfgang von Goethe, and is considered a masterpiece of nineteenth-century literature. Six incomplete English translations of *Faust* appeared not long after its publication, one of which is the verse translation published anonymously by Thomas Boosey in 1821. Attempts have been made to attribute this translation to Samuel Taylor Coleridge, though the attribution remains controversial. The present discussion tests the hypothesis that Coleridge was its author.

The general approach of the discussion that follows is stylometric. The rapidly growing availability of digital electronic literary texts since the mid-twentieth century offers an opportunity to supplement traditional literary-critical techniques with mathematically and statistically based computational methods in literary analysis, and the academic discipline devoted to development of this methodology has come to be known as stylometry. The motivation for adopting a stylometric approach here is that its analytical results have the fundamental scientific properties of objectivity and replicability: they are objective in the sense that they are based on mathematical and statistical methods which are generic to data analysis rather than application-specific, and replicable in that anyone with access to the data and analytical methods used to generate the results can repeat and thereby confirm them.

Authorship attribution is a branch of stylometry whose remit, as its name indicates, is determination of the authorship of texts of unknown or disputed authorial provenance. This is done by comparing stylistic characteristics of texts of known authorship to those of the texts to be attributed, where style is defined in terms of features identified in data abstracted from text using mathematical and / or statistical methods. The focus of the present discussion is authorship attribution, and the class of methods selected to carry out the analysis is one that has thus far been relatively little used in the discipline: cluster analysis. Cluster analysis has long been used across a wide range of science and engineering disciplines as a methodology for discovering structure in data which is too complex for reliable interpretation by direct human inspection. Specifically, given a collection of objects described by some arbitrary, and typically large, number of variables which describe the objects, cluster analysis identifies the relative degrees of similarity among the objects on the basis of their respective variable values and represents the

similarity structure of the collection in an intuitively-interpretable graphical format. The motivation for selecting cluster analysis in the present application is discussed in detail in a subsequent chapter, but in essence it is its effectiveness in analysing data which describes text in terms of large numbers of variables.

This thesis comprises five chapters. The first chapter reviews the history and current state of the debate on the question of Coleridge's authorship of the 1821 *Faust* translation. The second states the research question being addressed and outlines the methodology used to address it. The third abstracts data from relevant digital texts, cluster analyses them, validates the analyses, and presents the analytical results obtained from the various clustering analyses. The fourth interprets the results of the analyses conducted in chapter three in terms of the research question. The five and final chapter summarizes and concludes the discussion. The conclusion is that, relative to the stylistic criteria and analytical methodology used, the proposition that Coleridge was the author of the 1821 Boosey translation of *Faust* is falsified.

The software used in the course of discussion to implement the analytical methodology described in chapter 2 is listed in an Appendix. This listing includes the code for several programs developed specifically for the present research application.

Chapter One

Motivation, History and Current State of the 1821 *Faustus* Translation Authorship Debate

This chapter reviews the history and current state of the debate on the authorship of the English translation of Goethe's *Faust* published by Thomas Boosey in 1821. It is divided into four main parts. The first part is my motivation for choosing this work to analyze. The second is a bibliographical overview of translations of *Faust* into English in the early 19th century. The third part reviews existing attempts to attribute Boosey's 1821 translation to Coleridge, and literary critical reactions to those attempts. The fourth part assesses the arguments for and against the attribution.

1.1 Motivation:

In 2007, Oxford University Press published a book entitled *Faustus from the German of Goethe translated by Samuel Taylor Coleridge* edited by Frederick Burwick and James McKusick, who presented evidence that the translator of an 1821 anonymous English translation of selections from Part I of Goethe's *Faust* was Samuel Taylor Coleridge. This book has been much debated and the stylometric analysis has been called into question by many reviewers, of which more will be said in due course.

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.

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 thesis.

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. In the stylometric section (2007: 330), McKusick admits that “the stylometric methodology presented here does not enable a persuasive answer,…” and 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. Details follow in the next chapter.

The scope of my empirical approach is extensive. I have examined not only Coleridge’s and other likely candidates’ involvement in the translation of *Faust*, that is, Staël, Soane, Anster, Boileau, and Gower, but also some other authors of the nineteenth century, namely, Wordsworth, Shelley, and Byron. The aim is to examine Coleridge’s literary style relative to the styles of contemporary authors to see where *Faustus* fits among them.

1.2 Bibliographic overview of the English translations of the first part of Goethe’s *Faustus* in the early 19th century:

Goethe published his *Faust*, the first part of the drama, in 1808. In 1809, Germaine de Staël undertook a translation of various scenes from *Faust* into French. Staël’s presented her translation in *De l’Allemagne* (On Germany) which was published in Paris in 1810 (Constantine, 2006, 2005; Classe, 2000; Hauhart, 1909; Haney, 1902; Boyle, 1987).

Like the English translations of Schiller’s dramas, the Staël translation of *Faust* attracted considerable publishing interest. Publishers of English translations of German’s literature particularly John Murray, Thomas Boosey, and Johann Heinrich Bohte, as will be discussed below, decided to translate and publish the play and make extracts from of it available to English readers.

Staël's *De l'Allemagne* was first published in England by John Murray in 1813 in the original French. It was subsequently published by John Murray that year in an anonymous English translation (Fitzsimmons, 2008; Constantine, 2006; Classe, 2000; Hauhart, 1909; Haney, 1902; Boyle, 1987; Smiles, 1891).

Recently, however, Burwick and McKusick (2007) suggest that the anonymous translation of this edition is by Francis Hodgson. According to Burwick and McKusick (2007), the title-page of Staël's 1813 edition after the words "translated from the French" is marked by the following pencil annotation: "by Francis Hodgson ed. by Wm. Lamb". Burwick and McKusick take this as evidence to attribute Staël's extracts of *Faust* to Hodgson. For more information on the Hodgson's translation of Staël's extracts of *Faust*, see Burwick and McKusick (2007: xvi, 114-117). On the other hand, Murray (2009a:3) in his review article of Burwick and McKusick's 2007 book doubts that Burwick and McKusick are correct to assert that Hodgson is the translator.

In 1815, Percy Bysshe Shelley attempted his own translation of part of *Faust* (probably to practice his German), which was literal, almost word-for-word and contained many errors (Constantine, 2006, 2005; Reiman, 2002, 1977; Stokoe, 1926). In the following years, as he improved his German, Shelley successfully translated two scenes from *Faust*: "Walpurgis-Night" and the "Prologue in Heaven", which appeared in 1822 in volume I of *The Liberal* (O'Neill and Howe, 2013; Reiman, 2002, 1977; Fritz, 1971; Marshall, 1960; Mary Shelley, 1824).

Another translator is George Soane, whose first translation of extracts of *Faust* was published by a German bookseller in London, Johann Heinrich Bohte, in 1820 (Glass, 2005; Hauhart, 1909; Reiman, 1977). Soane re-attempted a translation of *Faust* in 1821 for Bohte as well, but for some reason, he only completed lines 1-576, i.e. roughly one third of the play (Glass, 2005; Reiman, 1977; Fritz, 1971; Hauhart, 1909). Four years later, Soane substantively reworked Goethe's text and his reworked translation appeared in 1825 in *Faustus: A Romantic Drama* (Mays, 2012).

The next translator is John Anster. He translated lines 1-1600 for H. Bohte in 1820 using blank verse to avoid Goethe's difficult words and phrases (Dowden, 2011; Fitzsimmons, 2008; Casey, 1981). His translation, *Faust: A Dramatic Mystery* appeared in *Blackwood's Edinburgh Magazine* in 1820 and received considerable attention from the reviewers. For

example, it was described variously as “closely imitating Goethe’s varied verse” (Hauhart, 1909: 124), an “adaptation...that has changed the content of the poem and has distorted the characters of Faustus and Gretchen” (Classe, 2000:596), “a brilliant paraphrase”, and “an almost incredible dilution of the original” (Bayard Taylor, 1871: 357).

In 1820, the London publisher Thomas Boosey published an anonymous partial translation of *Faust* with illustrations (Fitzsimmons, 2008; Fritz, 1971; Hauhart, 1909). Recently, Burwick, (2008a) and McKusick and Burwick (2007: xix) reveal that the translator of this edition was brought out under the pseudonym “a German in humble circumstances”, who is found to be Daniel Boileau. Again, in his review article, Murray (2009a:3) sees no reason to think that Daniel Boileau translated this edition of *Faust*, arguing that Burwick and McKusick attributed it to him with no evidence.

One year later, Thomas Boosey undertook a second translation of *Faust* and published it anonymously in 1821. This edition included most of Part I and was translated in verse and connected by a prose narrative (Fitzsimmons, 2008; Fritz, 1971; Hauhart, 1909). According to Burwick and McKusick (2007), this translation is the work of Coleridge and the short title “STC Faustus 1821” is used and repeated regularly throughout their 2007 book (Burwick, 2008a; Burwick and McKusick, 2007).

Finally, Lord Francis Leveson-Gower made a translation of *Faust* which was published by John Murray in 1823. In the preface to this edition, Gower admitted that his knowledge of German was inadequate and did not deny that he did not attempt to translate several parts of Goethe’s text because of the difficulties he encountered in keeping the original meaning in the translation (Hauhart, 1909: 99).

1.3 Existing attributions of Boosey’s 1821 *Faustus* to Coleridge:

The 1821 Boosey translation has been variously attributed to Francis Hodgson, 1813 (Staël’s translator according to Burwick and McKusick), George Soane (1820, 1821, and 1825), John Anster (1820), Daniel Boileau (1820), Leveson Gower (1823), and, recently, 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*.

Nothing was said on the subject until 1971 when Paul Zall, a scholar of English Romanticism and American literature, used traditional stylistic analysis, that is, qualitative authorship attribution, to argue for the attribution to Coleridge (Burwick and McKusick, 2007). Zall's methodology was simple: he looked for stylistic similarities between works known to be by Coleridge (the translation of *Wallenstein* and *Piccolomini*, his plays *Remorse* and *Zapolya*) and the 1821 *Faust* translation. Based on his analysis, Zall stated that there were 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.

On this basis, Zall assumed that Coleridge was the actual author of the *Faust* translation and that he published his work anonymously in 1821: "...the lost work was perhaps never missing at all, but merely disguised under the cloak of anonymity...if it is not by Coleridge then there was an imitator at large who deserves better of posterity than unsung anonymity...." (Grovier, 2008: 2; Shimek, 2007). 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. Zall commented that "they just simply wouldn't believe that Coleridge translated Faust...there were many rejections, and finally I said, 'to hell with it, life is too short', so I switched over to other things" (Burrowes, 2007: 1). In 1989, however, Zall passed the materials along to Jim McKusick, who reviewed them.

After 15 more years, in 2003, Frederick Burwick joined McKusick to re-examine Zall's materials with much greater detail. The two scholars make their case that Coleridge was the author. This case is articulated in a book titled: *Faustus from the German of Goethe Translated by Samuel Taylor Coleridge* and published by Oxford University Press in 2007 (Mays, 2012, Burwick, 2008a and 2008b, Burwick and McKusick 2007, Shimek, 2007).

Burwick and McKusick's case is based on three types of argument: (i) circumstantial historical evidence, (ii) qualitative stylistic criteria, and (iii) quantitative stylistic criteria, that is, stylometry. These arguments, together with the ones advanced by various literary scholars and the present discussion, are considered separately in what follows.

1.3.1 The circumstantial historical argument:

Burwick's historical argument (Burwick and McKusick, 2007: xv-xxxv) relies mainly on external evidence such as biographical documentary record using notebooks, conversations, as well as incidents and circumstances in Coleridge's life and works and events in the composition of the 1821 *Faustus* translation to connect Coleridge to a *Faust* translation. The argument also relies on a series of letters between Murray and Coleridge, Boosey and Coleridge, Boosey and Goethe, and Bohte and Goethe.

This section summarizes this evidence which Burwick presents for Coleridge authoring the 1821 *Faustus* translation.

First, Burwick claims that Coleridge is involved in the translation of Goethe's *Faust*, "not once but twice". (Burwick, 2008:4; Burwick and McKusick, 2007: xxx). The first time was in 1814; John Murray (the London publisher) asked Coleridge to do the job. Though the wages were regarded by Coleridge as "humiliatingly low", he nevertheless accepted and signed a contract. Murray gave him £100. After working on the translation for two and a half months, probably on the grounds that the play offended Coleridge's Christian views, he changed his mind and the contract was broken; Coleridge could not produce the translation, nor could he return the money to Murray (Burwick, 2008a; Burwick and McKusick, 2007:xxiv).

The second time was in 1820; when Burwick claims Coleridge translated *Faust* for Thomas Boosey (Murray's serious rival publisher). According to Burwick, the most likely sequence of events would go something like this. In May 1820, Boosey planned another publication, with additional scenes from *Faust*, to go with the second edition of Moritz Retzsch accompanied by twenty seven plates engraved by Henry Moses. Boosey was looking for a qualified translator to do the translation *Faust*. He asked Coleridge for "friendly advice". Coleridge thought Boosey was asking him whether, as translator, he would do the translation himself. Coleridge told Boosey that he was willing to do it if he would be given the right to explain the play's moral and religious issues in another way. Coleridge also suggested that a blank verse drama mixed with prose summaries would be the best way to translate and represent Goethe's text. After negotiating the offer, Boosey and Coleridge agreed on the terms. Burwick provides Coleridge's detailed plan entitled "My Advice and Scheme" dated 12 May 1820, which contained correspondence between Coleridge and Boosey, including Coleridge's reply to Boosey "friendly request", to

support his claim for Coleridge's involvement. Further, Coleridge insisted Boosey to keep his name concealed and the whole project be published anonymously. Coleridge's letter to Boosey of 10 May 1820 makes it evident that he requested his identity as the translator of *Faust* to be kept unknown "...without my name I should feel the objections and difficulty greatly diminished..." (Letters 5:42-44). However, after months, Coleridge finished the translation in September 1821 and Boosey published the work with no mention of the translator's name (2007: xix- xxi). Burwick confirms that Boosey preserved the translator's anonymity in the announcement in the *London magazine* of July 1821: "the publishers of Moses's Etching from Retzsch's Outlines to the Faustus, have engaged 'a Gentleman of literary eminence' to prepare a translation of a considerable portion of that wild and singular play into English Blank verse" (*London Magazine*, 1821: 104; "works preparing for publication" cited in Burwick, 2008a:4).

Second, Burwick (Burwick and McKusick, 2007: xxiv) claims that Coleridge demanded his identity to remain anonymous for at least three reasons:

- (i) According to Coleridge's opinion, much of *Faust's* language was "blasphemous", "vulgar", and "licentious" (Boyle and Guthrie, 2002:145; Hauhart, 1909: 65; 69). The text also contained themes and questionings of religion that made Coleridge uncomfortable.
- (ii) He wished his identity to remain unknown in order not to undermine his reputation through a partial translation that showed him unable to bring it to a finished state.
- (iii) His unfulfilled previous commitment to John Murray and his fear that Murray would pursue him for the £100 he owed.

Related to the above, McKusick, as cited in Murray (2009a:4) and Shimek (2007), speculates the situation upon which Coleridge's agreement with Boosey was reached: "it went something like this: Coleridge said, 'Yes, if you pay me, I can produce a verse translation quickly-- because it's almost done-- but you must swear never to reveal my name as the translator. It must go to the grave. Otherwise, Murray will come after me for his 100 pounds, plus interest, plus breach of contract.'"

Third, Burwick (Burwick and McKusick, 2007) presents a letter that Boosey's rival publisher, J. H. Bohte sent to Goethe on 1 August 1820 telling him that Coleridge was working on his *Faust* for Boosey. In this letter, Bohte wrote: "under the progressive

cultivation of German literature in this country one has become especially attentive to your *Faust* to which the splendid outline engravings by Retzsch have contributed much... I hear with pleasure that the poet Coleridge is working on a complete translation of this Dramatic poem". Burwick and McKusick (2007: xxi) and Burwick (2008a:4) say that the letter is a "smoking gun" that would serve as important evidence supporting the case for Coleridge as the true translator of the play. Burwick (Burwick and McKusick, 2007: xv) presents also another letter dated 4 September 1820 from Goethe to his son August, as a response to Bohte's letter, repeating his news that Coleridge was translating *Faust* (Burwick, 2008a:4; Burwick and McKusick, 2007: xv, xxi; Burrowes, 2007).

Fourth, since there was an exchange of letters between Boosey and Goethe and Bohte and Goethe, Burwick (Burwick and McKusick, 2007: li) believes that Boosey and Bohte were the source of information to Goethe, that is, they informed him that "the English *Faust* of 1821" was the translation of Coleridge. He also believes that once Goethe received this information, he, therefore, on 8 May 1826 in his diary, pointed out to Coleridge's connection to his *Faust* "Antheil von Coleridge" ('Coleridge's part') and to Boosey's edition that contained the Retzsch's plates "Kupfer von Retsch zu *Faust* nachgestochen" ('Retzsch's copperplates for *Faustus* reproduced').

Fifth, Burwick provides an answer to or explanation for Coleridge's famous statement "I never put pen to paper as translator of '*Faust*'" (*Table Talk*, 1833) by saying that evidence from Coleridge's letters and the rumour of his circle of friends reflects his efforts at translating Goethe's *Faust* on two occasions and this constitutes a conclusion that Coleridge did "not only put pen to paper", but "had done so with ardour and determination". (Burwick and McKusick, 2007: xxx, Burwick, 2008a:4). McKusick, in respect to Coleridge's denial as well, as cited in Saut Ste. Marie (2007), contends that "He lied...He was covering his own tail".

Finally, from all of this, Burwick concludes that there are no sufficient grounds by which to suspect Coleridge's authorship of *Faustus*. For anyone in Coleridge's circle of friends, the translation appears to say such a thing. Though this is known to a few, "to Boosey, to Anster, to Goethe, and no doubt to the Gillmans and a few others in the Coleridge circle, the fact of Coleridge's translation was gradually forgotten".

The final piece of evidence pointing to Coleridge's involvement in the translation is found

by Burwick and McKusick in the preface to William Barnard Clarke's translation of *Faust* parts I and II in 1865. In this edition, Clarke refers to Coleridge as the translator saying that an earlier translation "said to be by Coleridge" (Burwick and McKusick, 2007: liv).

1.3.2 The qualitative stylistic argument:

The most important qualitative stylistic features, or what Burwick terms "verbal echoes" from Coleridge's other works repeated throughout different scenes of the translation of *Faust*, are as follows:

- (i) Coleridge's blank verse is rarely characterized by end-stopping lines and occasionally characterized by the use of a preposition or adjective at the end of a line, which prompt both sense and sound forward into the next line like II.3240-5 from "Forest and Cavern" scene (2007: xxxv) (Burwick and McKusick, 2007: xxxvi):

There may I gaze upon
The still moon wandering through the pathless heaven;
While on the rocky ramparts, from the damp
Moist bushes, rise the forms of ages past
In silvery majesty, and moderate
The too wild luxury of silent thought.

- (ii) Coleridge's habit in repeating certain phrasal patterns is present in the translation of *Faust*. This occurs in his earlier works: "fancy's wild hopes" in *Remorse* (1813), "thy heart's wild impulse only dost thou..." in *The Death of Wallenstein* (1800), "...endearment, All sacrificed to liberty's wild riot" in *The Fall of Robespierre* (1794), "...young-eyed Joys! Advance! By Time's wild harp" in 'Ode to the Departing Year' (1796), "...of vernal Grace/And Joy's wild gleams that lighten'd o'er..." in 'Monody on the Death of Chatterton' (1790). Such repetitions also occurred in different parts of the 1821 *Faust* translation – for example, "my soul's wild warfare...", "my heart's wild tempest..." (2007: xiii).

- (iii) Mephistopheles's monologue on the ascent of the Brocken in the 1821 *Faust* translation, which deviated from reliance on blank verse, echoes the rhythmical power of the four metrical feet of 'Christabel' (Burwick, 2008a:1; Burwick and

McKusick, 2007: xxiii). This is clear from one of the reviews on the Boosey edition Burwick refers to being found in the *European Magazine* published in October 1821, where the reviewer cited the ascent of the Brocken and described it as equivalent to Coleridge's 'Christabel' saying: "There is a wild rush in the above lines, which at once make them very life they describe; they come to the ear like the night blast over a bleak hill...yet it is surely the work of which no man ashamed. Rumour says the author of Christabelle tried at it and resigned it" (Burwick and McKusick, 2007: xxiii).

(iv) In 1827, Coleridge wrote to James Gilman: "...we have had and have a steady deliberate soft thick soaking Rain, which yet does not sufficiently disburthen the Atmosphere of its ever contracting and dilating, ascending and descending aqueous vapor, as to quiet the gusty winds or to smooth the white breakers..." (Letters 6.706). In this letter, Coleridge used the same words and phrases that Goethe used in the original text to describe an image in the Sign of the Macrocosm: "Golden buckets, like the paddles of a water-wheel, are seen as scooping up the heavenly powers and forever ascending and descending as the wheel resolves (Goethe's *Faust* 499). (Burwick, 2008a:1).

(v) According to Burwick (Burwick and McKusick, 2007: xxxviii), Coleridge's habit of changing the meaning from the original source text in translation is recognizable in the following lines from the translation of *Faustus*:

How divinely
Are all things blended how each lives and moves
But with the rest how heav'nly powers descend
And re ascend balancing reeling worlds...

Here, according to this view, Coleridge changed the meaning and deviated from Goethe's original images and words to those of his own poetic idiom, a skill which reflects the characteristics of Coleridge's descriptive style (Burwick and McKusick, 2007: xxviii; Burwick, 2008a:1).

(vi) In some of his other verse, Coleridge tends to use the same words that he wrote upon his own first ascent of the Brocken (the highest peak of the Hartz mountains) in the

countryside during the Hartz walking tour. For example, Coleridge used the effect of Hartz poetically in ‘Lines Written in the Album at Elbingerode’, in the Hartz Forest (1799):

Stood on Brocken’s Sovran height, and saw
Woods crowding upon woods, hills over hills,...

The same effect recurred in the translation of *Faustus* (in the image that arises from the scene of “A Forest and Cavern”) when Coleridge departed from Goethe’s original vague image to the one he himself developed by using images drawn from his own first climb of the Brocken in Germany (1798-9):

While on the rocky ramparts from the damp
Moist bushes rise the forms of ages past...

Here the phrase “rocky ramparts” parallels the phrase “proudly ramparted with rocks” that occurred in Coleridge’s ‘Ode to the Departing Year’ (1796) (Burwick and McKusick, 2007: xxxviii; Burwick, 2008a:2):

...Proudly ramparted with rocks
And Ocean mid his uproar wild...

(vii) Phrases such as “the forms of other days” and “the forms of Memory” from Coleridge’s ‘Anna and Harland’ (1790):

For fair, tho' faint, the forms of Memory gleam,...

or “the faded forms of past Delight” from Coleridge’s ‘To Robert Southey’ (1795):

Thy sadder strains, that bid in MEM'RY's Dream
The faded forms of past Delight arise;...

are also parallels to the phrasing that recurred in Boosey’s text:

Moist bushes rise the forms of ages past
In silvery majesty and moderate...

Here one point must be made for the credibility of the current discussion. Burwick (Burwick and McKusick, 2007: xxxviii) makes an error with the parallel cited in Coleridge's 'Anna and Harland' above. After we examined the content of this poem in the original source text of Coleridge's poetical works (e.g. Mays, 2001, part I: 27; Coleridge, 1912: 17), the mistake is evident as the line in this poem says: "The tales of other days before me glide:..." not "the forms of other days...".

- (viii) The use of the word "witchery" in a two-word phrase, or what Burwick terms "Coleridge's habit to empower witchery with a participle" (Burwick and McKusick, 2007: xlv), is another parallel that occurred (only once as we examined it) in the *Faust* translation: "the soul with juggling witchery...". This feature also occurred in Coleridge's other works: "it mocks my soul with charming witchery" in *Piccolomini* (1800), "soothing witcheries" in 'Songs of the Pixies' (1793) and "floating witchery" in 'The Eolian Harp' (1795).
- (ix) The phrase "Silent thought", which occurred in Coleridge's *Biographia Literaria* (1817) "... in your mind...Such stores as silent thought can bring", also repeated in Boosey's text: "...The too wild luxury of silent thought..." (Burwick and McKusick, 2007: xxxvi).
- (x) According to Burwick, Coleridge's hand in the translation of *Faust* is clear, not only in terms of the verbal patterns but also in terms of the ideas and images that he employed in his previous works (Burwick and McKusick, 2007: xxxvi). Attention may be called to the phrase "great spirit". Coleridge used this phrase with the "tone of devotion" in his early sonnet 'To William Lisle Bowles' (1794) with the "tone of devotion": "Like that great Spirit, who with plastic sweep...Mov'd on the darkness of the formless Deep!" or in the prayer of Alvar in *Remorse* (1813): "kneeling I prayed to the great Spirit that made me...". The same phrase with the "tone of devotional thanksgiving" is recognized in Coleridge's translation of *Faust*: "Oh, thou great Spirit, thou hast given to me...All, all that I desire. Thou hast not turned..." (Burwick and McKusick, 2007: xxxvii).
- (xi) Phrases such "bright hopes", "enlight'ning dull", "no sweet imagining", and "To Nature", "Beam on my darkling spirit" had no equivalents to stand in Goethe's text; i.e. they are, as Burwick and McKusick state, Coleridge's addition to the text.

(Burwick and McKusick, 2007: xxxix-xl).

Burwick goes on to say that the 1821 *Faust* translation echoes words and phrases characteristic of Coleridge's earlier works of 1814-20: about 10 percent of the vocabulary is peculiar to *Remorse* and *Zapolya*, and certain other words are peculiar to poems written about 1820 (Burwick and McKusick, 2007: xliv).

1.3.3 The quantitative stylistic argument:

McKusick's role was to find quantitative evidence in support of the joint claim of Coleridgean authorship (2007: 312-30). To this end, he compiled a digital electronic corpus comprising:

- (i) Four plays by Coleridge: *Remorse* (1813) and *Zapolya* (1817) written by him, and *The Death of Wallenstein* (1800) and the *Piccolomini* (1800) which he translated, as already noted.
- (ii) The anonymous Boosey 1821 translation of *Faust*.
- (iii) Five other translations of *Faust* by Hodgson (1813) Staël (1809), Soane (1821 and 1825), Anster (1820), Boileau (1820), and Gower (1823).

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*; examples for *Remorse*, Anster's translation of *Faust*, and for Gower's translation of *Faust* are given in Figures (1.1), (1.2), and (1.3) respectively. An explanation of why they are reproduced is deferred to Chapter Four in order not to pre-empt the discussion. For the moment, it is enough to see how much each work is similar to/or different from *Faustus*.

STYLOMETRIC ANALYSIS

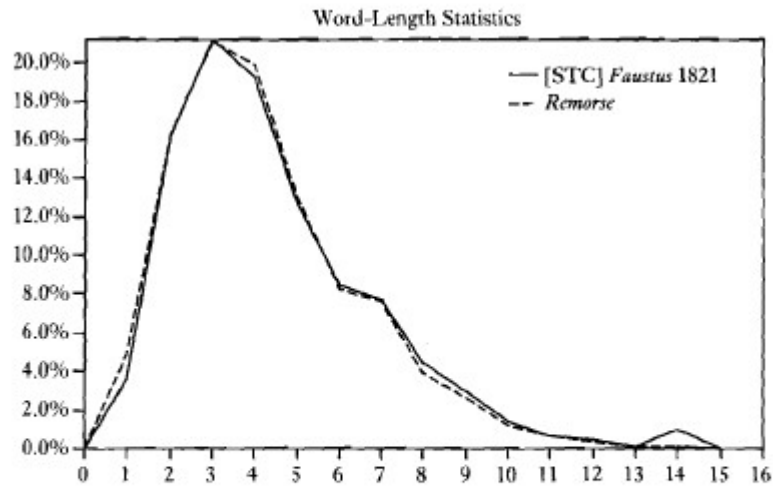


Fig. 1

Word-Length Statistics

Text Name	Total	1	2	3	4	5	6	7	8	9
[STC] <i>Faustus</i> 1821	16859	3.6%	16%	21%	19%	13%	8.5%	7.7%	4.5%	2.9%
<i>Remorse</i>	13048	4.8%	16%	21%	20%	13%	8.3%	7.6%	3.9%	2.6%

Chi-Square significance test (columns 2-8): Chi-Square value = 7.824 Chi-Square 20% value : p-value = 0.2512 In this case the difference is *not* significant even at the 20% level.

Figure (1.1) word length measurement for *Faust* 1821 and *Remorse* taken from McKusick and Burwick (2007:317)

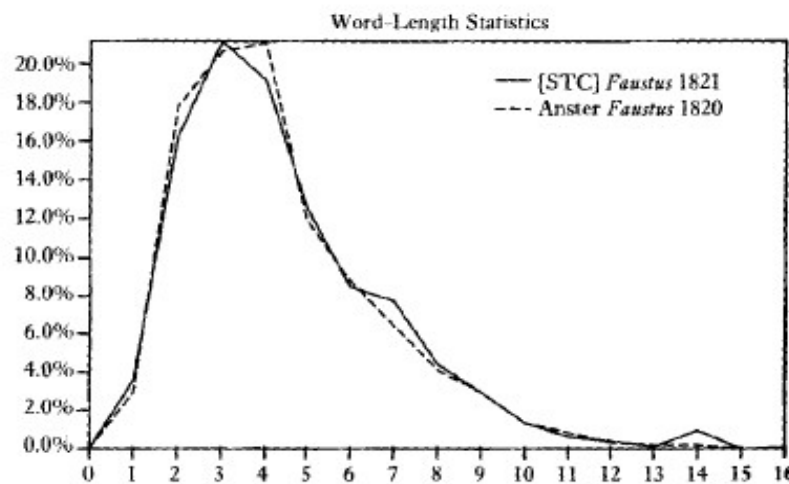


Fig. 2

Word-Length Statistics

Text Name	Total	1	2	3	4	5	6	7	8	9
[STC] <i>Faustus</i> 1821	16859	3.6%	16%	21%	19%	13%	8.5%	7.7%	4.5%	2.9%
Anster <i>Faustus</i> 1820	16053	3%	18%	21%	21%	12%	8.8%	6.5%	4.2%	3%

Chi-Square significance test (columns 2-8): Chi-Square value = 51.127 Chi-Square 0.1% value p-value < .0001 In this case the difference is very highly significant, at the 0.1% level.

Figure (1.2) word length measurement for *Faust* 1821 and Anster's translation taken from McKusick and Burwick (2007:318)

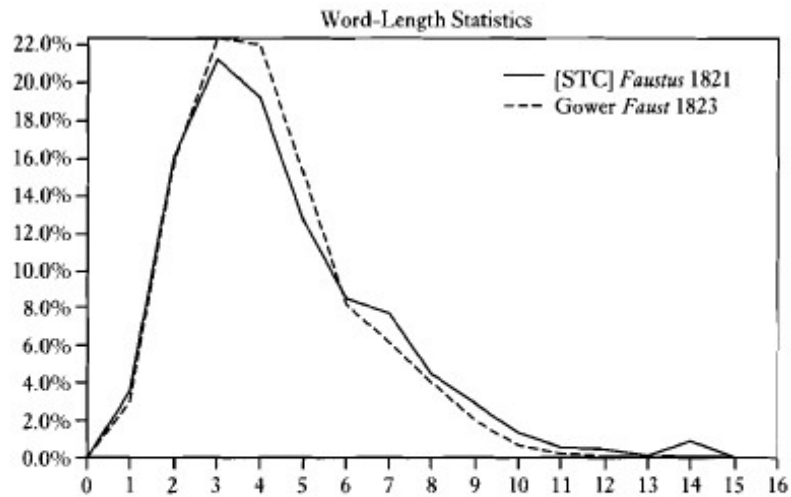


Fig. 4
Word-Length Statistics

Text Name	Total	1	2	3	4	5	6	7	8	9
[STC] <i>Faustus</i> 1821	16859	3.6%	16%	21%	19%	13%	8.5%	7.7%	4.5%	2.9%
Gower <i>Faust</i> 1823	3384	3%	16%	22%	22%	15%	8.2%	6.2%	4%	2%

 Chi-Square significance test (columns 2–8): Chi-Square value = 31.586 Chi-Square 0.1% value = p-value < .0001 In this case the difference is very highly significant, at the 0.1% level.

Figure (1.3) word length measurement for *Faust* 1821 and Gower’s translation taken from McKusick and Burwick (2007:317)

He then applied the chi-squared test (e.g. Balakrishnan et al., 2013, Greenwood and Nikulin, 1996, Shrirama, 2005) 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 word-length 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).

1.4 Assessment:

Literary critical reaction to Burwick and McKusick’s claims has been mixed. Reviewers include, Mays (2012), Engell (2012), Uhlig (2010), Murray (2009a, 2009b), Schmid (2009), Paulin et al. (2008), Crick (2008), Craig (2008a), Grovier (2008), Fenton (2008), Lomenzo (2008), Bode (2008), Robertson (2008), Carlyle (2008), and Burrowes (2007), and there are rebuttals by Burwick (2008a, 2008b, 2010). The positions taken by these respondents vary considerably, ranging from those who see the claim for Coleridgean authorship of the 1821 translation as entirely lacking strong evidence and based too much on conjecture, through to those who see some merit in the claim, to those who are convinced by the evidence which Burwick and McKusick offer. This section argues that while Burwick and McKusick have not made a convincing case for the attribution of the 1821 *Faust* translation to Coleridge, the various responses to their work cited above do not successfully refute it.

1.4.1 The present discussion’s own reaction:

Why Coleridge? In general terms, he seems a reasonable candidate on account of his competence in the German language and his interest in German literature. Coleridge studied the German language and acquired considerable knowledge of German literature before his departure to Germany in 1798. In a letter to Thomas Poole in May 1796 Coleridge wrote: “...on very trivial and on metaphysical subjects I can talk tolerably...I can read old German, and even the old low German better than most of even the educated

natives...” and “chief efforts were directed towards a grounded knowledge of the German language and literature...and in about six weeks I shall be able to read that language with tolerable fluency...”(Turnbull, 1903: 81; Haney, 1902: 5 Coleridge, 1834: 122).

While he mastered the German language, Coleridge wrote a small number of original poems in German and translated some other poems into German. He also adapted and translated a large number of poems from various German poets such as Schiller, Lessing, Voss, Wieland, and Goethe into English. Just to name a few of them, in 1796 Coleridge translated his sonnet ‘To a friend, who asked how I felt, when the nurse first presented my infant to me’ into German (Mays, 2001:369); in 1797 he translated ‘The Wieland’s Oberon’ (Mays, 2001:540); in 1798 he adapted ‘English Duodecasyllables’ from Matthisson (Mays, 2001:694); in 1799 he wrote ‘Hexameters’ (Mays, 2001:696) and imitated ‘Hymn to the Earth’ from Stolberg’s ‘Hymne an die Erde’ (Mays, 2001:617) and ‘Tell’s Birth-place’ from Stolberg’s ‘Bei Wilhelm Tells Geburtsstatts im Kanton Uri’ (Mays, 2001:624); in 1799 he translated a passage in Ottfried’s ‘Metrical paraphrase of the Gospel’ (Coleridge, 1912:304); in 1798 and 1799 he translated and imitated from Schiller ‘The Homeric Hexameter’, ‘The Ovidian Elegiac Meter’, ‘The visit of Gods’ (Mays, 2001:699, ‘A Distich’ (Mays,2001:1050), and ‘Ossian’ (Mays, 2001:735), and translated ‘Epigrams from Lessing’ (Mays, 2001:792). For more on these works, see J.C.C. Mays (2001).

More importantly, Coleridge decided to translate all the works of Schiller, and once the first editions of Schiller’s *Wallenstein* (a drama in two parts) arrived in England, Coleridge translated both *Piccolomini* and *The Death of Wallenstein* into English. Coleridge’s translations were a great success. So the reviewers, J. G. Lockhart, William Wordsworth, John Hookham Frere, Lady Caroline Lamb, and Ludwig Tieck, praised him in the printed publications of the time for the elegance of his translation, as quoted by Leigh Hunt in the *Literary Examiner* in 22 November, 1823 and in the *London Magazine* in August, 1824 (Mays, 2001).

After his success in translating Schiller’s works, Coleridge’s confidence and ability in the translation became so well being aware of German language and German life, and, therefore, he showed himself able to repeat such a success in dramatic translation.

Obviously, this explains why contemporaries such as Henry Crabb Robinson and P. B. Shelley held the belief that only Coleridge at that time was able to produce a good

translation of *Faust*. Henry Crabb Robinson said of him that "...there is no doubt that Coleridge's mind is much more German than English... he is eminently qualified to bring the literature of Germany to the attention of his countrymen" (Hauhart, 1909: 63) and P. B. Shelley was convinced that "no one but Coleridge is capable of this work" (i.e. translating the whole of *Faust*) (Mays, 2012:123; Murray, 2009:2; Constantine 2006: 221; Hauhart, 1909: 95). This also explains why the publisher John Murray in 1814 entered into negotiations with Coleridge and tried to convince him to translate it for him or the publisher Thomas Boosey in 1820 asked Coleridge for "friendly advice" or for doing the translation. As part of his mastery of German, two literary scholars alluded to the possibility of Coleridge having translated *Faust*. One of them is William Hauhart (1909: 32, 95), who shows little doubt that Coleridge's translation might be among the other translations that failed to come to light. The second is Rosemary D. Ashton (1977:156-67), who states that almost everything indicates the possibility of Coleridge's missing translation; she comes to the conclusion that *Faust* was to be one of the many of Coleridgean projects that "never got off the ground". Still today many scholars continue to support this possibility as will be discussed below.

The circumstantial historical argument for Coleridge's authorship of the 1821 *Faustus* is based on two main pieces of evidence. One is essentially that Coleridge had a good knowledge of German, was interested in Germany and its literature, had done translations into English of other German literary works, and had contracted to do a translation of Goethe's *Faust*. This makes him a strong candidate for authorship of the 1821 *Faustus*, but logically does no more than that. Another is essentially that there was an exchange of letters between Boosey and Coleridge and between Boosey and Goethe which are (according to Burwick and McKusick) the strongest links to the *Faust* translation. But should they be taken as a reason to attribute the *Faust* translation to Coleridge?

Based on the biographical documentary record associated with the early British reception of *Faust*, nothing in these letters confirms the involvement of Coleridge in a *Faust* translation. There is no record that Coleridge ever claims involvement in the translation of *Faust* during his life time, neither publically nor secretly, in person or in writing, other than his famous denial in *Table Talk* in 1833 "I need not tell you that I never put pen to paper as a translator of Faust".

We believe, like many others, that if Coleridge was the translator, there should be

definitive evidence of it from his life time, from himself or from his circle of friends or Highgate household or even from his literary enemies. There is none. Not that there are no reasons to think that Coleridge translated *Faust*, but we find them inconclusive; additional evidence is required.

The qualitative argument claims to provide evidence that Coleridge is actually Boosey's translator. Burwick and McKusick's argument relies on verbal parallels and similarities of vocabulary usages occurred in Coleridge's works in drama and poetry and in *Faust* translation. They conclude that "over 800 verbal echoes..." and "echoes, combined with the cadence and metaphorical texture of the blank verse, persistently reveal Coleridge as the translator" (Burwick, 2008a:1; Burwick and McKusick ,2007: xliv). To which the most we can say is that the existence of these types of parallels and similarities of words and phrases do not automatically mean there is any connection between Coleridge's works and the *Faust* translation. Why?

There are two justifications. First, similarities between any two authors' writing styles are inevitable. This is also particularly true of the Romantic period, as in case of any literary period (e.g. Elizabethan era), where it is always possible to find similarities of words and phrases between any two authors not because those words or phrases are unique to either author, but simply because they are the convention, or even clichés, of the time. The related literature of the Romantic period shows that many authors, who were connected by the ideas and philosophies they shared and experimented with artistic forms and styles, consciously or unconsciously, borrowed words, phrases, and aesthetic images from each other and also from other writers (e.g. some authors influenced by Shakespeare, imitated and borrowed from him) and used them to describe things in their own works according to the language habits, or, more technically, the norm of the time (Murray, 2009 and 2004; Turley, 2009; Mazzeo, 2006). Here is a list of just a few examples to support the discussion. Coleridge, Southey, and Wordsworth borrowed words from each other and used them in their own works in the 1790s (Murray, 2004:202), Keats and Cornwall borrowed freely from each other when each wrote his sonnet 'Bright Star' in 1819 or 1820 (Turley, 2009; Murray, 2009:7), Shelley's 1816 poem 'Mont Blanc' deliberately echoed Coleridge's 'Hymn before Sunrise in the Vale of Chamonix' and knowingly engaged with its themes and symbols, Yeats borrowed the use of contrasts (good vs evil-young vs old, art vs nature, body vs soul) from Blake (Bornstein, 2006, 1970 as cited in "W.B. Yeats and the Romantics", Boston College Library, 2011); Byron borrowed from

Wordsworth (Mazzeo, 2006); Coleridge borrowed from Amos Cottle's translation of 'Edda of Saemund' (1797) the description of moonbeam "like April hoar-frost spread" and used it in his 'The Rime of The Ancient Mariner', and, in another position, he converted several words from Dante's 'inferno' (1308) into the stanza (lines 445-51) of that poem (Murray, 2007:7; 2004: 202). The literary borrowing of Romantic-period authors is discussed in such works as, for example, Mazzeo (2006). Even Burwick and McKusick themselves (2007: xxxiv) admit that Anster's poem 'The Times' imitated from Coleridge's 'France: An Ode and Reflections on Having left a place of Retirement'.

Even if it happens that in some cases a few distinctive similarities of words and phrases are found in any two authors or texts that still do not prove anything. The problem with using them as an attribution argument is that it really is not possible to be sure whether they occurred because they are by the same writer, or because they are—whether consciously or unconsciously— simply borrowed or imitated by writer X from writer Y.

What we needed from Burwick was some indication of frequencies (or the total number of occurrences) of these vocabularies and parallels or a comparison of these features with other writers of the Romantic period, or both. This might have provided a more convincing basis for the claim that these words and phrases were not found elsewhere in the writings of the same literary period as one might ask: how often has a given parallel or vocabulary occurred in Coleridge's works and *Faust*? Has the occurrence appeared in Coleridge's works and the *Faust* translation only, or it appeared in both and also once or more times in the five other translations by other suspect authors?

In other words, in order to identify a given word or phrase as stylistically unique to an individual author, we need first to define the norm of that author's own time (e.g. shared words or phrases, unique words or phrases, rare words and phrases, words or phrases with distinct meanings, characteristic words or phrases) then we assess the value of their occurrence against that norm (e.g. we may list and order all the features claimed to be idiosyncratic on the basis of their frequency of emergence) (Crystal, 1970:101-4, 1965:174-6, 1987:200-16, 1991:221-38, 1972: 103-14).

Secondly, words, phrases, rhythm, and concepts used in parallels can contribute to arguments about attribution but cannot be conclusive. A list of parallels of any two authors in any literary period, or especially in the Romantic period, could be compiled. In endorsing the validity of verbal parallels for authorship attribution, the literature (e.g.

Oakes, 2014; Love, 2002; Vickers, 2002; Criag, 1999; McMenamin, 1993; Bailey, 1979; Lake, 1975; Ashley, 1968; Baker, 1945; Sampley, 1933; Byrne, 1932; Oliphant, 1929) notes that there are dangers in relying on parallels of word, phrase, rhythm, and concepts, such as ones that Burwick presented as evidence for authorship. These have long been looked at very sceptically by literary critics or reviewers for several reasons. The most relevant ones are as follows:

i) Parallels can be assigned to authors by:

- a) Conscious or unconscious plagiarism
- b) Imitation, deliberate or otherwise
- c) Coincidence
- d) Convention or common literary resources

ii) Quality is all-important and parallels need to be graded with care. Not all parallels present in the texts are equally important, nor do they constitute “evidence”. Some may be frequent enough and unique to an individual author (I call them strong parallels); others may be rare and obvious elsewhere in a given literary period (I call them weak parallels). The key point here is about a parallel that dependent on an individual author rather than shared by writers of the same time period and genre, i.e. an idiosyncratic feature.

iii) Collecting a hundred “ungraded” parallels does not prove anything. The existence of parallels of words and phrases alone in the texts under consideration is not sufficient (i.e. they are likely to be of less value when used by themselves). However, combined with other parallels (e.g. of thought, rhythm, rhyme, or images), the detail would be much stronger but even that can be misleading. Many of those parallels can emerge coincidentally or from the language habits of the time.

The current discussion does not wish to suggest that similarities of words and phrases or parallels are valueless for authorship attribution. On the whole they are important and of value if they are selected cautiously with knowledge of the language habits of an author and the norms of the time, and if they are applied reasonably to disputed authorship problems. If a given feature appears in more texts, say, in the writings of three or four different authors it fails to have any value for authorship attribution, that is, whether the

feature is frequent or rare across the texts (Love, 2002; Crystal, 1987:200-16; 1970:101-4; 1965:174-6; Lake, 1975).

In summary, the qualitative argument of Burwick and McKusick (2007) is insufficient to connect Coleridge to the *Faust* translation.

The quantitative stylometric argument supports the case for Coleridge's authorship of the 1821 *Faustus*, but only weakly. As the review of stylometric methodology in the next chapter will show, 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 word-length distribution, between *Remorse* and the 1821 *Faustus*" is "suggestive". Function word distribution is a much better stylistic criterion, as will also emerge in the next chapter, but McKusick again claims only that it does not "prove" Coleridgean 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 Coleridgean authorship in this instance. He is thus right in claiming only that his results are "consistent with" the hypothesis of Coleridgean authorship, but his further claim that they "indicate a strong likelihood" of it is unjustified.

Overall, therefore, the view of the present discussion is that Burwick and McKusick go 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 hypothesis of Coleridge's authorship.

1.4.2 Other reactions:

Critical reaction to Burwick and McKusick's claims will be considered under the same three headings as above.

1.4.2.1 The circumstantial historical argument:

For the circumstantial historical argument, more biographical documentary materials that illustrate the early British reception of *Faust* are presented by Paulin, et al. (2008) arguing against the attribution. Many literary scholars such as Mays (2012), Murray (2009a), Robertson (2008), Lomenzo (2008), Hamilton (2008), Grovier (2008), and Crick (2008) agree with Paulin, et al. (2008); they share the very same points of views and draw similar conclusions. The general methodology directing the current study is empirical by nature, since it uses stylometry to test the hypothesis of Coleridge's authorship of *Faust*. As such, biographical documentary materials are not very close to what is required here and therefore is not concerned us. These can be found in Paulin, et al. (2008) 'A Gentleman of Literary Eminence'. However, for the balance of the current chapter, and because the arguments advanced by those scholars are in essence very similar to one another, a brief summary of these is given.

For those reviewers, Burwick and McKusick's historical evidence is insufficient and problematic to advance the case for Coleridge for the following reasons:

- (i) Burwick (Burwick and McKusick, 2007: xxxi) does not provide Coleridge's letter to Boosey dated 10 May 1820 with its attached note from "My Advice and Scheme" in full, nor does he trace Boosey's reply to it. However, this letter with its accompanying note from " My Advice and Scheme" and also Boosey's reply to Coleridge's letter, which fail to point to Coleridge's involvement in the translation, are available in full in Paulin, et al. (2008:5-7).
- (ii) The sequence of events is very obvious: Boosey knew that Coleridge had once planned to translate *Faust* for Murray in 1814 but changed his mind. At that time, Boosey was already formulating his plans for translating a new *Faust* edition and was looking for a qualified translator. Boosey asked Coleridge to do the task, made an offer of payment, and the two discussed the publishing agreement. Coleridge refused the offer but instead gave Boosey some poetical and text-translation advice. Boosey

thanked him and started looking for another translator following Coleridge's advice.

- (iii) During his lifetime, and from the late record of his *Table Talk*, Coleridge was asked whether he involved in the translation of Goethe's *Faust* on three occasions. One was in May 1825 with Giaocchino de Prati. Another was in his *Table Talk* on 16 February 1833, the year prior to his death, when he openly admitted: "I need not tell you, that I never put pen to paper as translator of Faust". A third was in his conversation recorded by John Hookham Frere in 1824. In each occasion, Coleridge consistently admitted that he "never put pen to paper as translator of Faust"

- (iv) Bohte's letter to Goethe dated 1 August 1820 is a common type of letter found in the publishing business. Bohte, as a leading German bookseller in London, was keen to keep Goethe informed of recent news and literary gossip. Importantly, the origin of Goethe's letter to his son August dated 4 September 1820, is completely without foundation and the letter itself is not known to have existed. Paulin et al., (2009:11) believe that this letter "is a speculation invented by Mckusick and Burwick needed to complete the conjectured series of events".

All in all, the conclusion of the counter-historical argument is essentially that there is no direct historical record that Coleridge actually translated *Faust*, nor any extant letter or reference that he was involved in its translation.

1.4.2.2 The qualitative stylistic argument:

Reviewers look at Burwick's internal evidence by focusing on two areas: stylistic similarities and the quality of the translation of *Faustus* itself. However, these two areas have divided reviewers in their reactions to Burwick and McKusick's edition.

(i) Reviewers against the attribution:

- a. Similarities of style. A list of words, expressions, and phrasal patterns that were suggested by Burwick and McKusick as parallels and verbal echoes from Coleridge's other poetry occurred in the translation of *Faustus* are examined. Reviewers argue that these are not unique to Coleridge's writing style and would therefore point to a number of authors. In other words, they can all be found anywhere else throughout the Romantic-era literature, details of which can be found in Murray (2009a: 6-8, 2009b)

and Crick (2008:78-9, 84).

- b. The quality of the translation of *Faustus*. The play is critically examined and evaluated in terms of its aesthetic forms and manner of expression. Several passages and verses from different scenes of the play are examined and compared with Coleridge's translating style. For literary scholars, the 1821 translation of *Faustus* is not consistent with Coleridge's previous translated dramas, arguing that its stylistics does not conform to Coleridge's style. This suggests that the anonymous Boosey translator of *Faust* was lacking in artistic skills appropriate to English poetry. Details of this can be found in Engell (2010), Murray (2009a), Guido Kohlbecher as cited in Murray (2009a:9-12), Scott (2008), Crick (2008:77-8, 80-1), Robertson (2008: 248-50).

(ii) Reviewers for the attribution (fully or partially):

- a. Similarities of style. Crick (2008:78-9) and Grovier (2008) believe that there are some passages and lines that clearly revealed some of Coleridge's idiosyncratic stylistic features, but, at the same time there are also some that are not unique to Coleridge at all and are found elsewhere. Schmid (2009-1-3) believes that Burwick's internal evidence is persuasive and that the many verbal echoes of Coleridge's style found in the Boosey's text revealed his unacknowledged familiarity with Goethe's *Faust* drama. Burrowes (2007), Shimek (2007), and Bode (2008) believe that the verbal echoes and phrases in Boosey's text occurred in Coleridge's other works, for example in *Remorse* are sufficient in constituting proof that Coleridge actually translated *Faust*, describing them as extensive and beyond random coincidence.
- b. The quality of the translation. Crick (2008:80-1) believes that in his translation of the pair of Schiller dramas, Coleridge used far more prose in more scenes than Schiller did, and that the same happened in the translation of Boosey's *Faust* text, but with less technical ability in the English prose. More specifically, the translation of Boosey's text required a translating strategy or practical experience, of which Coleridge was fully aware. This strategy was clearly explained when Coleridge wrote to the publisher Murray in 1814: "A large proportion of the work cannot be rendered in blank verse, but must be in wild lyrical meters". Robertson (2008: 248-50) and Engell (2010) state that the excellence of movement in some translated passages in blank verse from Boosey's text made them wonder if Coleridge might have involved in its

production. Schmid 2009 (1-3), who shows full agreement with the attribution (Murray, 2009a:2), believes that in comparison with the five other translations, Coleridge's translation of *Faust* in 1821 is the best in quality, not only because of his poetic skills but also because of his extensive knowledge of German literature, theology, and philosophy.

1.4.2.3 The quantitative stylistic argument:

Like the circumstantial and qualitative stylistic pieces of evidence, McKusick's stylometric evidence (2007: 312-330) also faces critical hostility (Murray, 2009a, 2009b; Paulin et al. 2008; Crick, 2008; Craig, 2007) for eight main reasons:

- (i) Based on the standards of the scientific methodology related to the presentation of the statistical results, Paulin et al. (2008:4) and Crick (2008:70) argue that the title page of the 2007 edition—*Faustus: From the German of Goethe Translated by Samuel Taylor Coleridge*— was not “framed as a question nor yet as a hypothesis” but as a fact that it was indeed Coleridge who translated Boosey text and published it anonymously in 1821.
- (ii) In terms of statistical significance, the results should be presented with confidence and accuracy. As Crick (2008: 71) indicates, McKusick presented the results of his statistical analysis with care “suggest a strong likelihood that Coleridge was the translator of the 1821”. However, based on this, Craig (2008: 8); Murray (2009a: 8-9), and Paulin et al. (2008: 27-28) point out to three main limitations found in McKusick's presentation of the statistical results:
 - a. The weak statistical information obtained from the comparison between Boosey's *Faust* and Coleridge's *Remorse*, and more limited comparisons with the five other translations of *Faust* indicates in the first place a conclusion which is “fully consistent with the hypothesis” (pp 325, 327). Furthermore, this conclusion becomes a “strong statistical correlation” (pp xlv, 312). In other words, weak evidence led to strong conclusions.
 - b. The contradiction in the presentation of results is clear: in one place “stylometry deals in probability, not certainty” (p.327), in another “this kind of analysis is no longer considered definitive or even particularly reliable by stylometrists” and finally

McKusick suggests that “it is nevertheless possible to gain interesting and suggestive results by looking at this kind of data” (p.313).

This contradiction in the statements of the results, according to Craig (2007:8), suggests that McKusick was not sufficiently sure of the accuracy of the analysis.

c. The statistical analysis is limited to two stylometric variables: counting the relative frequency of word lengths and function words, which McKusick makes too much of. Word-length frequency distribution cannot be relied on (to distinguish between authors) for attribution argument; it has been judged to no longer be informative about authorial style. For the result of an attribution to have a clear meaning, the statistical values cannot consist of limited measurements. McKusick must have examined all the possible kinds of textual measurements available to him. Taking this position into consideration, it seems unhelpful for McKusick to put any weight on the results obtained from this measure. (Craig, 2008: 85-6).

(iii) For any stylometric analysis to be accurate, the sample size must be large enough to be appropriately measured. McKusick failed to provide a sufficiently large corpus.

(iv) McKusick’s corpus consisted of Coleridge’s four plays (*Wallenstein*, *Piccolomini*, *Remorse*, and *Zapolya*), as well as translations of *Faust* by other writers of the time, of course, including the 1821’s translation. However, the results of statistical analysis show that “there is a similarity in vocabulary, as reflected in word-length frequency distribution and the frequencies of ten function words, between *Remorse* and the 1821 *Faust*” (Burwick and McKusick, 200: 318-324), but this evidence does not conform with the standards for the use of stylometry in authorship attribution, which are as follows:

a. There are objective criteria one can apply to the comparison of texts to establish its reliability. Of this requirement, Murray (2009b:8) and Craig (2008: 87) argue that interpreting data and drawing any conclusions in terms of the similarity of one text of known attribution with another (of disputed authorship) would seem unreliable for determining authorship; any single text, particularly a short one, could reveal distinctive stylistic features for many different reasons. The analysis is far more reliable and the result is valid if it is conducted on a large number of texts.

b. McKusick compared *Remorse* to *Faust*, but he failed to provide results for the

comparisons between *Faust* and Coleridge's other plays (i.e. other texts are not taken into consideration). McKusick claimed that he performed these comparisons, but neglected them because the results did not support his argument: "there is a statistically significant difference between the 1821 *Faustus* and each of the other three Coleridge plays (*Wallenstein*, *Piccolomini and Zapoloya*)" (Murray, 2009a: 8; Craig, 2008: 86).

c. According to Murray (2009a: 9), the deliberate omissions of unwanted results are inexcusable for three reasons:

1. Each of the neglected plays has relevance to the 1821 *Faust*: of all Coleridge's plays, *Zapolya* (1817) is the text closest to *Faust*'s date, and the other two plays, *Wallenstein* and *Piccolomini* (1800), are also translated from German.

2. As discussed above, comparison of only a single text to another is insufficient.

3. According to the standards of scientific research, the results ought not to be neglected. Omitting results usually signals a clear failure to apply stylometric methods in a manner that will provide reliable results and also means that readers will not be able to assess and understand how statistically different the other texts are from *Faust*.

(v) The five other candidate translators tested in McKusick's stylometric attempt may not qualify as reasonably strong competitors. There are two reasons for this claim. The first is that it is not possible for a translator who had once translated one work to then translate it again quite differently. The second is that McKusick failed to give each of those translators the very same opportunity to show similarity to *Faust* as Coleridge. In the corpus there were four texts belong to Coleridge against a single text for each of the other five candidate translators (Craig, 2008: 86-87). McKusick probably thought that these were the only possible candidate authors and that a translation of the same text gave each candidate author set a good opportunity to show similarity, if any, to the 1821 Boosey's *Faust* that additional comparison was unnecessary. Related to this limitation, Crick (2008: 83) agrees with Craig (2008) in arguing that in terms of the statistical and stylistic analysis, the analysis was bias because more data was considered in relation to Coleridge than to other possible translators. This is also true in terms of the historical and qualitative arguments which considered Coleridge's known works (poetry, plays, translated plays, etc), relations with his publishers, his letters, and biographical interpretation of literary-historical records. No such attention

exists for any other five proposed alternative candidate translators.

- (vi) The result carries mixed evidence for Coleridge's authorship since the similarity of *Faust* to *Remorse* suggests authorship to Coleridge, while its difference from the other three Coleridge plays suggests the reverse. (Craig, 2008: 86).
- (vii) Aware of these limitations in the statistical analysis, Grovier (2008) believes that McKusick's results drawn from the analysis of the relative frequency of word-lengths and function words offer "preponderance of evidence" that left little doubt that Coleridge was involved in the translation of Boosey's text.
- (viii) Bode (2008), Shimek (2007) and Burrowes (2007) believe that there is no reason to doubt McKusick's evidence that the anonymous translator of 1821 Boosey's text was Coleridge. No such direct evidence exists for any other candidate translator: the statistical profile of Coleridge's features found in the translation of 1821 *Faust* was not reached by any other *Faust* translations. For them, the statistical evidence is conclusive and, as Shimek and Burrowes go further and exaggerate, is comparable to "Coleridge's fingerprints" (Burrowes, 2007) and "Coleridge's literary DNA" (Shimek, 2007) that found on *Faust* translation.

In conclusion, the arguments from both sides—those in favour of or against the attribution—fail to provide a conclusive demonstration (documentary or statistical) of a connection between Coleridge or any other candidate translator with the 1821 Boosey translation of *Faust*. All we can claim here is that there is room for reasonable doubt about the case for Coleridge's authorship of *Faust* as expressed by literary scholars, and that other reasonable scenarios are possible. In other words, the *Faust*-Coleridge debate remains open. In the light of this and of McKusick's invitation (Burwick and McKusick, 2007: 315-16, 327, 330) for stylometrists to pursue further analysis to examine the hypothesis made by himself, the current thesis broadens the scope of the quantitative investigation by including not only the six related *Faust* translations but also a large number of works belongs to Coleridge and a few romantic contemporaries, by considering a large number of variables, and by applying advanced methods in the domain of attributional stylometry.

Chapter Two will define the research question to be addressed together with a methodology for doing so. The methodology will then be applied in Chapter Three.

Chapter Two

Research Question and Methodology

2.1. Research question:

In their re-edition of Boosey's 1821 *Faustus* translation, Burwick & McKusick articulated the hypothesis that Coleridge was its author. The present discussion tests that hypothesis.

2.2 Methodology:

This section is in three parts:

- Part 1 outlines the nature of the authorship identification problem as it is understood in the current state of the discipline.
- Part 2 reviews the literature on authorship identification.
- Part 3 describes the methodology used by the present discussion to address the research question posed in (2.1) above.

2.2.1 The authorship identification problem:

Authorship identification has historically been part of the more general field of stylometry, whose aim is to augment the qualitative methods used in traditional philology and literary criticism for the study of text with theoretical tools and methodologies drawn on the one hand from linguistics and on the other from mathematics and statistics; overviews of the field are given in (e.g. Oakes, 2014, 2002; Grzybeck, 2014, 2007; Bruce et al. 2012; Stamatatos, 2009, 2008; Koppel et al. 2009, 2002; Shlomo, 2008; Juola, 2008; Forsyth, 2007; Grieve 2007, 2002; Nieto, 2004; Baayen et al., 2002; McEnery and Oakes, 2000; Holmes, 1994, 1995, 1989, 1998; Baayen, 1996). As its name implies, the aim of the subdiscipline is to identify the authorship of text where this is disputed or unknown. The literature has identified the following classes of authorship identification problems.

- *Closed-class problem:*

This is also known as the multiple authorship or n -class problem, where $n \geq 2$ (Juola,

2008; Binongo, 2003; Diederich et al., 2003; Fung, 2003; Juola & Baayen, 2003; Holmes et al., 2001; Baayen et al., 2002). It addresses a situation in which there is an anonymous or disputed text and a set of writers who are thought to be reasonable candidates for authorship of it. Sample texts from the candidate authors are studied to determine the characteristic style of each, and these characteristic styles are compared to that of the text of interest to determine which of the candidates is the most likely author. Where n is large, this type of problem is also known as the needle-in-a-haystack problem: Madigan et al. (2005) have considered 114 authors, Luycks & Daelemans (2008) 145, and Koppel et al. (2002, 2006, 2012) thousands of candidates.

- *Open or one-class problem:*

This is also known as authorship verification (Juola, 2008; Koppel & Schler, 2004), and differs from the closed-class problem in that it involves only one candidate author: given disputed text and a corpus of work by that author, the aim is to decide whether he or she wrote the disputed text.

- *Profiling authorship problem:*

It is also known as the characterization problem. In this case there is no candidate set of authors. Instead, the task is to derive evidence from the style of a given text about its author, such as the writer's age, gender, ethnicity, and so on (Juola, 2008; Koppel et al., 2002; Stamatatos, 1991).

2.2.2 Literature review:

Stylometrists (e.g. Oakes, 2014; Kestement, 2014; Stamatatos, 2009; Koppel et al., 2009; Juola, 2008; Argamon and Levitan, 2005; Binongo, 2003; Peng et al., 2003; McMenamin, 2002; and Holmes, 1998, 1994, 1985; Bailey, 1979) generally assume that one part of an author's writing style is conscious, deliberate, and open to imitation or borrowing by others. The other is sub-conscious, that is, independent of an author's direct control, and is far less open to imitation or borrowing. Stylometry focuses on the unconscious part of an author's writing style and assumes that at least some aspects of it are constant across his or her literary output. Stylometrists further argue that these constants can be identified and applied to areas like authorship attribution on the basis of

quantitative criteria using computational methods.

The main foci in the development of stylometry have been (i) identification of unconscious stylistic features, called discriminators or variables, which can reliably be claimed to characterise the styles of individual authors and to distinguish them from the styles of others, and (ii) identification of specifically quantitative analytical methods which generate and use data derived on the basis of such variables in stylometric applications such as authorship attribution. The stylometric literature contains a large number of textual features suggested as discriminators of authorship (e.g. Grieve, 2002, 2007; Diederich et al. 2003; Juola and Baayen, 2003; Baayen et al. 2002, 1996; Kukushkina et al. 2001; Holmes et al. 2001; Dale et al. 2000; Stamatatos et al. 1999; Holmes, 1998, 1994, 1989, 1989; 1985) and quantitative analytical methods (e.g. Dabagh, 2010; Nieto, 2004; Koppel, et al. 2002; McEnery and Oakes, 2000; Hair et al. 1995; Holmes, 1994, 1998). This section surveys the subset of the literature specific to authorship attribution, which is the topic of the present discussion. It begins with a brief overview of earlier work in the field and then focusses in greater detail on developments from about the mid-twentieth century to the present; general surveys of stylometry together with more detailed discussion of older work on authorship attribution are available in (Grzybeck, 2014, 2007; Bruce et al. 2012; Juola, 2008; Craig, 2008; Grieve, 2007, 2002; Forsyth, 2007; Koppel et al. 2002; McEnery and Oakes, 2000; Baayen, 1996; Holmes, 1998, 1995, 1994, 1989).

2.2.2.1 Older works:

The history of stylometry goes back to the work of Jewish scholars in antiquity, who attributed the *Torah* to Moses based on the analysis of the style and the structure of verses in the *Torah* and the subsequent books of the Old Testament. At that ancient period, two early practices of stylometry are identified: (i) counting of the number of verses, words, and letters in addition to the number of occurrences of certain words in each book of the Old Testament to ensure accuracy in transcription, and (ii) looking for hidden meanings in letter patterns and for the numbers that could be derived from them.

More recently, eighteenth and nineteenth-centuries Europe saw a growing interest in the problems of authorship attribution, notably for the purpose of identifying the authorship of older works such as the *Iliad* and the *Odyssey*, the different books of the *Bible*, and the

works of Shakespeare. In 1713, for example, Richard Bentley considered the question of whether the *Odyssey* was written by the same poet as the *Iliad*, concluding on the basis of stylistic features that a single poet composed the *Iliad* for male listeners and the *Odyssey* for women. In 1795 Heinrich Wolf argued, again on the basis of stylistic features, that the *Iliad* and the *Odyssey* were created before the invention of writing, and that the poems they contained must be regarded as a collection of songs or short stories that had originally composed one by one. In 1787 the Shakespearean scholar Edmond Malone argued that the three parts of *Henry VI* were not really written by Shakespeare, to whom they were traditionally attributed.

Perhaps the most influential contribution to the field of authorship attribution is that by the English mathematician Augustus de Morgan, who in 1851 gave new insights into how an authorship attribution problem of a given text can be solved. One of these insights, which related to the classical problem of the authorship of the biblical *Epistle to the Hebrews*, was to compare different-length words used in Greek text generally with those in the other Pauline epistles. To solve the problem of authorship, de Morgan suggested, in his own words, to “count a large number of words in Herodotus—say all the first book—and count all the letters; divide the second numbers by the first, giving the average number of letters to a word in that book...do the same with the second book. I should expect a very close approximation...” (Taken from de Morgan’s letter to his friend Rev. W. Heald as reproduced in his wife’s *Memoir of Augustus de Morgan*, 1882: 215-216 and cited in full in Unsworth, 2013).

Attempts to develop his quantitative method and to find new methods had continued by de Morgan himself in 1880 and by other researchers to examine an author’s a literary style up until 1965. Here are some famous attempts:

- Conrad Mascol (1887, 1888) used the relative frequency of punctuation marks, average sentence length, and the relative frequency of function words to examine the Pauline Epistles.
- Mendenhall (1887) examined Dickens’s word-length frequency distribution in *Oliver Twist* and Thackeray’s in *Vanity Fair*. He also examined (1901) the word-length frequency distribution for all works written by Shakespeare, Bacon, and Marlowe.
- Sherman (1888) introduced sentence-length frequency distributions as a way to characterize authors’ styles.

- Lutoslawski (1897) used stylistic elements to establish a chronology for Plato's various dialogues. This involves the fact that it was Lutoslawski who first introduced the term “stylometry” in 1890 and defined its general principles as a group of methods for “measuring stylistic affinities”.
- Thorndike (1901) introduced the use of contractions as a style marker for determining the relative contributions of Shakespeare and Fletcher to the jointly authored play *Henry VIII*. This methodology was subsequently used by Farnham (1916) to examine several works by other Elizabethan authors.
- In 1939 Yule used sentence length statistics to examine various works attributed to Francis Bacon, Samuel Taylor Coleridge, and Charles Lamb. Five years later, he used the same statistics to examine the authorship of *De Imitatione Christi* and *Bills of Morality*. Further, Yule developed another statistics, called “Characteristic K” or “Yule K”, to calculate word repetition rates irrespective of text length. Yule used this measure to examine the relative numbers of nouns occurring once, twice, and so on in a number of texts and published the results in his *The Statistical Study of Literary Vocabulary* in 1944.
- In 1949, E. H. Simpson developed a statistics to characterize an author’s style by measuring the probability of occurrences of arbitrarily chosen lexical words. He called his statistics the “Simpson's D”, which is closely similar to Yule’s K.
- In 1949 Sir William Elderton used word length measure to characterize an author’s number of syllables per word.
- In a series of publications (Fuchs 1952, 1954; Fuchs and Lauter 1965), the mathematician Wilhelm Fuchs examined average word length in syllables, word-length frequency distribution in syllables (i.e. the ratio of word tokens with one syllable, the ratio of word tokens with two syllables, and so on), and the average distance between *n*-syllable words (i.e. the average distance between two one syllable word tokens) as distinguishing features between texts.

(Roper, et al., 2012; Juola, 2008; Grzybeck, 2007; Hockey, 2004; Nieto, 2004; Love, 2002; Grieve, 2002; Holmes, 1998, 1989; Rudman, 1998).

2.2.2.2 Recent developments:

The appearance and widespread diffusion of information technology in the second half of the twentieth century rendered the digital representation of text together with the abstraction and analysis of data from digital text readily practicable, and as a result

stylometry has developed rapidly. As noted earlier, developments in stylometric authorship attribution have focussed on the one hand on identification of suitable textual criteria for attribution, and on the other on development of effective quantitative methods for analysis of data based on such criteria. These are described in what follows.

a. Textual criteria:

For any of the following textual criteria, a fundamental distinction must be made between types and tokens. In domains such as logic, philosophy, science, computer, etc. the type/token distinction is a distinction that isolates a descriptive concept from objects that represent or embody the concept, considered as particular examples of it (see, e.g., Stanford Encyclopaedia of Philosophy and Wikipedia Type–token distinction). This distinction in the domain of stylometry and textual processing between types and tokens is similar and is used to determine the presence of a token, or types of token, and an occurrence of it. To understand the significance of this and distinguish between the two terms, it is necessary to consider the following example:

A rose is a rose is a rose

If we count the number of words in this sentence we get a total of 8 words. The number of words in a given text is often referred to as the number of tokens. However, in this sentence, a number of these tokens are repeated and there are only 3 different types. The token ‘*a*’ occurs 3 times, the token ‘*rose*’ occurs three times, and the token ‘*is*’ occurs two times. Tokens, therefore, are the total number of words or the occurrences of word types (i.e. they are particular concrete instances) and Types are the different words (i.e. they are unique and abstract). So, for the sentence in this example, there are 8 different tokens or occurrences of word types: 3 occurrences of the word type ‘*a*’, 3 occurrences of the word type ‘*rose*’, and 2 occurrences of the word type ‘*is*’. These are shown in Table (2.1).

Word	Token	Type
A	3	1
Rose	3	1
Is	2	1
Total	8	3

Table (2.1) An example of Type/Token ratio

This distinction is applied in a well-known measure, the Type/Token Ratio as will be seen in the course of discussion.

i. Word length:

The length of a word, defined as the number of letters which constitute it, is extensively used in stylometric authorship attribution because it is so easy to compute (Chaski, 2005): just count the letters. This measure is commonly applied in two ways: average word-length and word-length frequency distributions.

- Given some text T of interest, average word length is calculated by dividing the total number of letters in T by the total number of words in T .
- A word length frequency distribution for T is generated by first defining a sequence of lengths $L = 1, 2, \dots, n$ for some n corresponding to a reasonable maximum word length, say 30. The number of words for each length $1, 2, \dots, n$ is then counted, and the n values so obtained are plotted in ascending order of i , with the horizontal axis representing i and the vertical axis representing frequency. The result is a lexical frequency plot for T . If each of the n values is divided by the total number of words in T before plotting, the result is a probability distribution which is isomorphic with the frequency plot but scale-independent.

Word-length approach has come under criticism regarding its application in authorship attribution studies. Assumptions and conclusions have been advanced by a number of scholars (e.g. Grzybek, 2007; Grzybek et al., 2005; Kelih et al., 2005; Grieve, 2002; Collinge, 1990; Smith, 1983, 1985; Williams, 1970) which have suggested that word-length is not a characteristic of an individual author's style and that it tends to be under too much conscious control of an author. The conclusion is that word length is more a discriminator of genre or register or languages than authorship of disputed texts. If we compare a number of texts written by different authors in the same literary genre and around the same literary period with one another, their word-length distributions may appear so identical that they seem to have been written by one author. Smith (1983), as cited in Holmes (1994:88), concludes that "Mendenhall's method now appears to be so unreliable that any serious student of authorship should discard it".

Examples of stylometric studies which used word length are Tanguy et al. (2011), Iqbal et

al. (2010), Brennan & Greenstadt (2009), McKusick & Burwick (2007), and Seletsky et al. (2007), Grieve (2007), Hirst & Feiguina (2007), McEnery & Oakes, 2000; Forsyth et al. (1999), Foster (1989), Smith (1983), Rothschild (1986), Radday (1970), Williams (1970), O'Donnell (1966), Mosteller & Wallace (1964), Brinegar (1963).

A variant of word length as described above is to count the number of syllables per word; calculation of averages and generation of distributions proceeds as before; Forsyth et al. (1999) used this feature to examine the authorship of the *Consolatio Ciceronis*. There are also other variants whereby word length can be defined, such as the number of phonemes per word, but these have not been extensively tested and appear to be unreliable or impractical (Grzybek, 2007).

ii. Sentence length:

Sentence length is defined as the number of components which comprise it. Most often this is the number of words in a sentence, where a word is defined as in the preceding section, but other components are possible, such as the number of letters, or syllables, or specified syntactic units. This measure is typically applied in the following ways in the literature: average number of words per sentence, sentence length frequency distribution based on word frequency, average number of letters or characters per sentence, and sentence length frequency distribution based on letter or character frequency (Grieve, 2007). These are calculated or generated in ways analogous to the methods described with respect to word length above, and so the details do not need to be repeated here.

Like word length approach, sentence length approach has also been critiqued and disputed by researchers who used or examined it. A study done by Alvar Ellegard (1963) showed that that this approach is not useful for characterizing the style of an author since “the variability within each author largely overlapped the variability between authors”. Another studies (e.g. Kjetsaa, 1979; Mosteller & Wallace, 1980; Smith, 1983; and Juola, 2006) which conducted to shed more light on this approach found that sentence length works less well for discriminating authors according to style and suggested that this approach can be more useful to differentiate between genres or registers or languages than a study of disputed authorship. However, very few studies reported good results by using this feature in authorship attribution. For example, Tallentire (1972), Kjetsaa et al. (1984), and Mannion and Dixon (2004) found that sentence length was able to identify an

author's style and distinguish between various texts of disputed authorship. These studies also showed that the frequency distribution of sentence-length worked better than the average sentence length per text. Further study by Kjetsaa (1978, 1997) showed that sentence length measure had little distinguishing power on its own, but was very useful when combined with other features. (Holmes 1994; Grieve, 2002; Luyckx 2004).

The biggest disadvantage of using sentence length in authorship attribution is that it is assumed to be consciously generated by an author and that a change of punctuation when writing and moving from one sentence to another in a text has an effect on it (i.e. sentence length can be easily affected by changing punctuation). (Grieve, 2002; McEnery & Oakes, 2000; Holmes, 1994). Examples of attribution studies that considered sentence-length measurement include Brennan & Greenstadt (2009), Seletsky et al. (2007); Hirst & Feiguina (2007), Grieve (2007), Mannion & Dixon (2004); Holmes (1994), Kenny (1986), Mosteller & Wallace (1980), Kjetsaa (1978, 1979), Radday (1970), Herdan (1960, 1965), Morton (1965), and Wake (1957).

iii. Contractions:

A contraction is a shortening of an orthographic representation of a morphological element, such as 'don't' for 'do not'. This criterion counts the number of contractions found in a text, the basic assumption being that any given pattern of usage is unique to a specific author. Again, average number of contractions per text and distribution of contraction usage can be generated in ways analogous to those described for word frequency. Examples of attribution studies that used contractions include Tanguy et al. (2011), Farnham (1916), and Thorndike (1901). However, this feature is not well understood as a criterion for author attribution.

iv. Character and Word n-grams:

A character n-gram is defined as a string of contiguous alphanumeric symbols, perhaps including also punctuation symbols. For example, the clause 'the child laughed', which consists of 15 letters, consists of 15 1-gram tokens (T, H, E, C, H, I, L, D, L, A, U, G, H, E, D), 14 2-gram tokens (TH, HE, EC, CH, HI, IL, LD, DL, LA, AU, UG, GH, HE, ED), 13 3-gram tokens (THE, HEC, ECH, CHI, HIL, ILD, LDL, DLA, LAU, AUG, UGH, GHE, HED) and so on; in general, a text that contains x characters will contain $x - (n - 1)$

n -gram tokens. A word n -gram is defined as a string of words, where each n -gram is composed of n words. For example, the sentence “it is a new nice car”, which consists of 6 words, consists of 5 word bi-grams “it-is” “is-a” “a-new” “new-nice” “nice-car” and 4 word tri-grams “it-is-a” “is-a-new” “a-new-nice” “new-nice-car”).

The relative frequency of n -gram tokens are calculated by dividing the frequency of a given n -gram token, e.g. it-is-a, in a text by the total number of 3-gram tokens.

In the associated literature there has been much research examining n -gram, character or word n -grams. N -grams are first used for author attribution by Bennett (1967), and subsequently, for example, by Kjell (1994), Forsyth & Holmes (1996), Soboroff et al.(1998), Grieve (2002), Khmelev & Tweedie (2002), Kukushkina et al. (2002), Clement & Sharp (2003), and Eder (2011). Soboroff et al.(1998), Khmelev & Tweedie (2002), and Kukushkina et al. (2002) reported that the frequencies of occurrence of n -grams are useful for identifying the style of an author since they are content-independent and easy to measure. In 2011, Eder examined and compared the effectiveness of several lexical features including the most frequent words, word bi-grams, word tri-grams, word tetra-grams, letter bi-grams, letter tri-grams, letter tetra-grams, letter penta-grams, letter hexa-grams, and different letter sequences in an attempt to identify which traceable features can be evidence of authorial characteristic of style. The results of this test showed, as reported, that letter n -grams are slightly less accurate than single words, and that word bi-grams and word tri-grams are generally useful for authorship attribution. For Forsyth and Holmes (1996) and Grieve (2007), word bi-grams and character n -grams are able to capture the style of specific authors better than lexical features. Dunning (1994:16), as cited in Luyckx (2004), reported very good results using n -grams and encouraged stylometrists to use this approach in authorship studies to attribute disputed texts. He argued that n -grams tends to work well for authorship attribution because they are similar to common words and short as well. Another studies by Koppel et al. (2009), Stamatatos, (2009), and Houvards & Stamatatos (2006) demonstrated that character n -grams are “sensitive to both the content and form of a text” and that character n -grams defined by a particular parameter n require a high-dimensional space to represent every possible combinations of words in a corpus (Stamatatos, 2013, 2009). However, the usefulness of n -grams is considered limited in authorship attribution, partly because many of character n -grams are closely related to particular “content words and roots” (Kestemont, 2014; Koppel et al., 2009), and partly because they require higher dimensionalities for their representation in space. This state of affairs leads to a potential problem in any given

application when texts have to be analysed and compared in terms of their distance (similarity or dissimilarity) from one another. Why? To represent a single word in terms of n parameter, many character n -grams are needed to capture enough stylistic or thematic information. For example, a single word such as ‘happy’ requires 5 1-gram tokens, 4 2-gram tokens, 3 3-grams tokens and so on for other words. A large character n -grams (say parameter n is 4 or 5 or 6) defines large pieces of stylistic and thematic information, some of them is redundant. The word ‘him’ requires 3 1-gram tokens, 2 2-grams tokens. A small n -gram (say parameter n is 2 or 3) doesn’t define or capture thematic information but still captures small pieces of, say, sub-word information such as syllable like information. How many possible n -gram types would be for higher than a 100.000-word corpus? The problem is that when the number of n -grams increases (character or word n -grams), the dimensionality increases greatly and the n -grams become increasingly sparse in the space they occupy, of which more will be said about dimensionality and sparse data in due course. Sanderson & Guenther (2006) and Coyotl-Morales et al. (2006) reported that the degree of accuracy performed by word n -grams is not always better than single or individual words.

v. Grapheme frequency:

This criterion measures the frequency of individual graphemes, that is, of individual alphanumeric characters, punctuation marks, or specialized symbols which a text might contain. For example, O’Donnell (1966) counted the frequency of dashes and semi-colons found in Stephen Crane’s unfinished novel *The O’Ruddy*, Chaski (2001) counted the frequency of a set of punctuation marks to examine and identify the distinctive punctuation habits of an author’s unedited texts, and Olsson (2006) examined stops, commas, question marks, exclamation marks, paragraphs, dashes, brackets, semi-colon, colons, and hyphens in connection with their syntactic roles in a very large corpus of texts. Others using this measure are Merriam (1988, 1994, 1998), Ledger & Merriam (1994), Ledger (1995), and Baayen et al. (2002).

In spite of the reported success in what is known as counting the frequency of graphemes in a text; for example, Iqbal et al. (2010) and Zhenshi (2013) argued that letter frequency and capital letter frequency were very reliable indicators of style, this approach is still unproven by researchers as a criterion for author attribution and even is discredited by some of them. For example, Love (2002), as cited in Grieve (2002: 26), criticised it

giving a particular reference to Merriam's 1994 attempt for not providing an explanation for using grapheme frequency as a discriminator of authorship. Forsyth & Holmes (1996) who assessed the usefulness of this approach reported that this feature is a poor criterion for authorship attribution.

vi. Vocabulary richness:

This criterion measures the degree of diversity of vocabulary in a text. It was introduced by Holmes (1985, 1989, 1994) as a reliable indicator of an author's characteristic style. Since then, use of vocabulary richness as a criterion has increased dramatically in the authorship attribution domain.

An obvious measure of vocabulary richness is the ratio of the number of word types to the number of word tokens in a text, commonly known as the type-token ratio V / N , where V is the vocabulary or number of types and N the number of tokens. This measure would appear to be independent of text length on account of the division by N , and would thus appear to make it possible to compare the type-token ratios of different-length texts meaningfully. This would in turn appear to make it possible to identify any given author's characteristic type-token ratio across his entire body of work irrespective of differing lengths of individual texts, and to compare that characteristic ratio of other authors' ratios, again irrespective of text length. Unfortunately, this has been found not to be valid. For any given author, the relationship between the number of word tokens in a text which that author generates and the number of word types it contains is in general nonlinear: in general, the number of word types grows at a slower rate than the number of word tokens, and so the type-token ratio for that author decreases as text length increases. In other words, the type-token ratio for any given author is not a constant, but rather it depends nonlinearly on text length (Stamatatos et al., 1999; Hoover, 2003).

To compensate for this effect, a variety of measures of vocabulary richness more complex than the simple type-token ratio have been proposed (Yule 1944; Simpson, 1949; Guiraud, 1954; Herdan, 1960, 1964; Mass, 1972; Honore, 1979; Sichel, 1975; Dugast, 1979; Holmes, 1985). Two frequently-used ones are described below.

- Yule's characteristic or Yule's K :

Yule's K is a complex measure for vocabulary richness of authors proposed by George

Yule in 1944. It is a measure of word repetition rates irrespective of its text length. The basic assumption behind the measure of this feature is that the occurrence of a given word is based on chance occurrence and can be understood as a Poisson distribution, that is, the number of times that a random and rare event occurs in some specified spatial or temporal interval. For more on Poisson distribution, see Clarke & Cooke (1998), Bell et al. (2009), and Holmes (1991).

However, this feature is calculated as:

$$K = 10^4 \frac{\sum r^2 V_r - N}{N^2}$$

Where 10^4 is an arbitrary constant used to avoid small and difficult to read K values, V_r represents the types or the number of different words used exactly r times (1, 2, 3 ...) in a text, and N represents the tokens or the length of text in words.

- Simpson's Index D :

This measure of vocabulary richness is related to Yule's K , and was proposed by E. H. Simpson in 1949 to measure the probability that two lexical tokens arbitrarily selected from a text will belong to the same type. This measure is calculated by:

$$D = \frac{\sum_i r(r-1)V_r}{N(N-1)}$$

where D represents the chance or probability, V_r represents the number of word-types that occur r times, for $r = 1, 2, 3, \dots, i$, and N represents the number of token- words in a corpus (Holmes, 1994).

These and other vocabulary richness measures have been applied to stylometric analysis by, for example, Guiraud (1954), Sichel (1975), Dugast (1979), and Miranda & Martin (2007). Having assessed these and other applications of the measures, Luyckx (2004) and Stamatatos (2006) concluded that they are unreliable when used in isolation as criteria for authorship attribution, but may be useful for corroboration when combined with other criteria. Tallentire (1972), as cited in Holmes (1994:93), reported that these measures are

ineffective to solve authorship attribution problems since the degree of word repetition or the occurrence of a vocabulary is probably under the conscious control of an author, that is, not a characteristic of writing style.

vii. Word frequency:

It has been claimed that word frequency, that is, the number of tokens of any given word type in a body of text, is a reliable criterion for authorship attribution (Kessler et al., 1997 and Karlgreen & Cutting, 1994), and, for this reason, it has been used in many attribution studies (e.g. Oakes, 2014; Koppel et al. 2009; Stamatatos, 2009; Grieve, 2007, 2002; Luyckx, 2004; Holmes, 1994; Dunning, 1994; Baayen, 2001; Binongo, 1994, etc.). This measure is easy to calculate, but selection of word types to calculate it for, that is, identification of which word types are the best indicators of authorial style, is problematic (Holmes & Forsyth, 1995).

The simplest approach to word type identification is to select those which an author uses most frequently, the assumption being that any given author has a characteristic preference for certain words and that the frequency of use of these words does not vary greatly across his or her literary output. Word frequency is consequently considered as a good criterion for identifying authorial style. Quite a few researchers have reported that this criterion successfully discriminates texts by different authors, for example, Chen et al. (2012), Dokow (2007), Grieve (2007), Madigan, et al. (2005), Stamatatos (2000, 2006), Luyckx (2004), Baayen et al. (1996), and Burrows (1987).

The reliability of word frequency as a criterion for authorship attribution is greatly improved by making a distinction between content words and function words. Content words are words with denotational semantics, and comprise the lexical classes of nouns, adjectives, verbs, and adverbs (Kula, 2010; Bell et al. 2009; Morrow, 1986; Clark & Clark, 1977). They are in general unsuitable for authorship attribution on account of the intuitively obvious observation that the choice of word types and their frequency of occurrence in any given text is topic dependent: an author writing about farming will select and frequently use different content words from one writing about astrophysics. Selection of content word types and their frequency of usage in a text are indicators of what the text is about, therefore, and not of authorial style. Since any given author may write on a variety of topics, and any number of authors may write on the same topic, the

unreliability of content words for author attribution is self-evident (e.g. Coyotl-Morales et al., 2006; Hoover, 2001, 2002, 2003a, 2003b). There may be particular circumstances under which content words are useful stylistic indicators--for example, an author's use of one or more very esoteric and therefore characteristic words--but in any specific application such circumstances have to be identified and justified.

More suitable as criteria for author attribution are function words, so called because their main linguistic role is to mark syntactic relations among content words: pronouns, auxiliary verbs, prepositions, conjunctions, determiners, degree adverbs, negations, quantifiers, and relativizers. Because of their primarily grammatical role, 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 than content words. Argamon & Levitan (2005), for example, showed that the frequencies of occurrences of different function words tend not to vary greatly across texts by the same author.

Many studies reported an increased use of function words in authorship attribution, for example, Kestemont (2014), Oakes (2014), Stamatatos (2009), Juola (2008), Argamon et al. (2007), Burwick & McKusick (2007), Bozkurt et al. (2007), Abbasi & Chen (2005), Zhao & Zobel (2005), Koppel & Schler (2003), Saric & Stein (2003), Juola & Baayen (2003), Binongo (2003), Fung & Mangasarian (2003), Baayen et al. (2002), de Vel et al. (2001), Holmes et al. (2001a and 2001b), Argamon et al. (1998), Kessler et al. (1997), Burrows & Craig (1994), Holmes (1994), Karlgren & Cutting (1994), Merriam & Mathews (1994), Burrows (1992), Morton (1978), Mosteller & Wallace (1964), there has been a few studies experimentally addressed the usefulness of function words as indicators of authorial style, of which more will be said in due course. Nevertheless, this approach has been criticized by some researchers (Hoover 2001; Oakes 1998; Oakman 1980; Damereau 1975; Tallentire 1972), mainly on the grounds that, because token frequency is dependent on text length, the derived function word frequencies have to be normalized relative to text length, but this normalization is unreliable for short texts (Moisl, 2008)

viii. Syntax:

This criterion assumes that each author has an unconscious characteristic syntactic usage

which distinguishes him or her from that of others (Stamatatos, 2009; Luyckx & Daelemans, 2005). Baayen et al. (1996) was the first study to propose syntactic features as a criterion for author attribution. They used the frequencies of occurrence of syntactic rewrite rules as an indicator of authorship. These frequencies were extracted from two syntactically annotated English corpora consisting of crime novels written by two different authors. The study found that the frequencies of occurrence of rewrite rules were able to distinguish between the texts of the two authors in question. In recent years, natural language processing (NLP) tools such as part-of-speech tagging and parsing have made it possible to use syntactic features of text as stylometric criteria: a set of syntactic features is selected, and the relative frequency of occurrence of these features across the texts being considered is then extracted. An example of syntax-based stylometric analysis is the phrase-level study by Stamatatos et al. (2001). This study measured the frequency of occurrence of various phrasal types--noun phrases, verb phrases, adverbial phrases, prepositional phrases, and combinations of these--in 300 Modern Greek newspaper articles. Stamatatos et al. (2001) reported that phrase-level features were “more robust for limited size of training data” and that “these features achieved higher accuracy than the lexically-based” ones used by researchers in authorship studies. However, this study failed to provide any discussion of validity enabling stylometrists to see whether there are any obvious problems or errors in the use of phrasal types to distinguish between the authors or articles tested. Instead, Stamatatos et al. (2001) concluded that “much else remains to be done as regards the explanation of the differences and the similarities between the authors”. More recently, a syntactic approach to authorship similar to that of Baayen et al.’s 1996 was adopted by Gamon (2004) who used a syntactic parser to measure re-write rule frequencies. The results of this study, as reported and cited in Stamatatos (2009:8), showed that the use of syntactic features alone to attribute authorship achieved bad results and that a combination of syntactic and lexical features can improve the results.

In fact there are two main problems with the use of syntax as a criterion for authorship attribution. The more important is the one it shares with all the other criteria already discussed: is the assumption that syntax is a reliable stylistic criterion justified? At present, not enough work has been done on this to support an answer (Juola 2008; Grieve, 2002; Stamatatos 1999, 2000, 2001; McEnery & Oakes 2000). The other is practical. Parsing and part-of-speech tagging technology have seen great improvements in reliability in recent years, but there is still a significant error rate, particularly for non-

standard and earlier forms of English and for other languages.

ix. Semantics:

A few studies have used semantic features of text as criteria for author attribution. For example, Deerwester et al. (1990) used lexical features to detect semantic similarities between words. Martindale and McKenzie (1995) and Craig (1999) examined the patterns of lexical choice in terms of the relative frequencies of content words. Hoover (2002, 2003) used sequences and collocations of content words. Gamon (2004) used binary semantic features (number and person of nouns, tense and aspect of verbs, and so on) and syntactic and semantic relations between a node of the graph and its daughters (e.g. a nominal node with a nominal modifier indicating location). McCarthy et al. (2006) extracted semantic features from synonyms and hypernyms, and Argamon et al. (2007) used a set of functional lexical features to represent the semantic function of each clause in a sentence and text (e.g. conjunction, elaboration, extension). None of the results reported in these studies, however, are particularly effective, however, and the use of semantic criteria for authorship attribution must therefore be regarded as requiring further development before they can reliably be applied. More recently, Tanguy et al. (2011), who used this approach and described it as rich stylometric features, concluded that simply using semantic features did not reach significant results. After all, semantic features have been proposed as a criterion for authorship attribution, but due to the complexity and relatively low accuracy of computational tools for semantic analysis, to use this approach, the results will not be accurate enough or are expected to have significant errors (Luyckx, 2010).

Here the researcher completes this section by showing 16 measurements for some of the stylometric features introduced above. Six lines from Coleridge's poem 'Virgin in a Roman Catholic village in Germany' (1811) are considered for this purpose. So in these lines:

Sleep, sweet babel! my cares beguiling:
 Mother sits beside thee smiling;
 Sleep, my darling, tenderly!
If thou sleep not, mother mourneth,
 Singing as her wheel she turneth:
 Come, soft slumber, balmily!

there are 31 words at 6 lines or sentences, 31 tokens, 27 types, and the following statistics:

- Average word length is 5.29 (calculation is made by dividing the total number of letters in the stanza (164) by the total number of words (31)).
- Word-length frequency distribution. In this stanza we find 7 words of length four, 5 words of length five, 4 words of length two, 4 words of length seven, 3 words of length eight, three words of length six, 2 words of length three, and 1 word of length nine. So, the frequency distribution of four letter-words is 0.22. (Calculation is made by dividing the total number of words of length (7) by the total number of words (31)).
- Average syllable count per word is 1.45 (Calculation is made by dividing the total number of syllables (45) by the total number of words (31)).
- The frequency distribution of words with 2 syllables is: 0.29 (Calculation is made by dividing the total number of words with 2 syllables by the total number of words (31)).
- Average sentence length per text is 5.16 (Calculation is made by dividing the total number of words (31) by the total number of sentences (6)).
- Sentence-length frequency distribution of 6 word-sentences is: 0.5 (Calculation is made by dividing the total 6-word sentences (3) by the total number of sentences (6)).
- Number of characters per sentence is 27.33 (Calculation is made by dividing the total number of characters (164) by the total number of sentences (6)).
- Average number of characters per word is 5.29 (Calculation is made by dividing the total number of letters in the stanza (164) by the total number of words (31)).
- Number of contractions per the stanza is 0.13 (Calculation is made by dividing the relative frequency of contractions (4) by the total number of words (31)). (The researcher assumed that there are four contractions in the six lines above).
- The stanza contains 24 hapax legomena and 2 hapax dislegomena.
- Type/Token ratio per text is $27/31 = 0.870 \times 100 = 87.09\%$ (Calculation is made by dividing the total number of types by the total number of tokens, and the ratio is multiplied by 100 to express it in percentage).
- Yule's K per text is $10^4 \times 6695/961 = 69.66$ high diversity (Calculation is made by multiplying 10000 by the sum of dividing the word types of each observed frequency to the power of two and the number of word types observed with that frequency by the total number of all tokens multiplied by its self).
- Simpson's D index is $10/31 \times 30 = 10/930 = 0.01$ high diversity (Calculation is made by dividing the sum frequencies of each type word by the product of multiplying the total

number of tokens by the total number of tokens minus one).

- CW/FW ratio is $21/31=76.74\%$ (Calculation is made by dividing content words to the total number of words and the resulted ratio is multiplied by 100 to express it in percentage).
- The relative frequency of function words is $9/31=0.29$ (Calculation is made by dividing the sum of the relative frequencies of all function words by total number of words).
- The relative frequency of content words is $21/31=0.68$ (Calculation is made by dividing the sum of the relative frequencies of all content words by total number of words).

b. Quantitative methods:

Whatever the stylistic criteria used to derive it from text, it had to be analysed in order to generate useful results. Several academic disciplines devoted to the application of quantitative and more specifically statistical methods to the analysis of natural language speech and text exist, including computational linguistics, corpus linguistics, natural language processing, and information retrieval, each of them with extensive literatures, and a review of these is out of the question here. Instead, what follows reviews the methods which have actually been used in stylometrics to date; more general information about work in related disciplines is available in (e.g. Moisl, 2015; Mirkin, 2013; Everitt et al., 2011; Hair et al., 2010; Berkhin & Dhillon, 2009; Gan, Ma, and Wu, 2007; Izenman, 2008; Gordon, 1999; Kaufman & Rousseeuw, 1990, Berkhin & Dhillon, 2009); briefer accounts are (e.g. Oakes, 2014, 2008; Koppel et. al., 2009; Juola, 2008; Grieve, 2002; McEnery & Oakes, 2000; Holmes, 1998; 1992).

Historically, attribution methods used in authorship attribution were statistical univariate methods measuring a single textual feature, such as example word length, sentence length, frequencies of letter n -grams, and distribution of words of a given length in syllables. Common univariate methods are T-test, which compares the averages of two samples and Z-score, which calculates the mean occurrence and the standard deviation of a particular feature and compares it within the normal distribution table, and these are covered in detail in the standard statistics textbooks, for example, (Woods et al., 1996).

These univariate methods were used to analyse texts in terms of a single stylometric

criterion or two and the results derived from them are therefore described as a simple form of statistical analysis. One of the most famous studies of this approach was carried out by Mendenhall T. C. in 1887. In this study, Mendenhall used histograms of word-length distribution to examine texts attributed to Bacon, Marlowe and Shakespeare. Many researchers followed Mendenhall's methodology from 1887 up to the present time. For example, Bringar (1963) examined the relative frequency distribution of word length in the disputed letters *Quintus Curtius Snodgrass*. Merriam (1993) used univariate Z-score to examine Shakespeare's plays, Marlowe's *Tamburlaine* and the disputed *Edward III*. One year later, Merriam (1994) used the count of letter frequency as a discriminator of authors in 43 plays. Burrows (2002) used univariate Z-score to examine a collection of texts by 25 authors from Restoration era. More recently, McKusick and Burwick (2007) examined the frequency distribution of word length in the disputed 1821 *Faustus* translation and several works by other suspect authors. Other examples of the univariate approach are those based in Bayesian probability and cumulative sums.

i. Bayesian Probability:

This is one of the earliest attribution methods, based on reasoning from the Bayes' theorem of probability. It was used in 1964 and 1984 by Mosteller & Wallace to examine a problem of disputed authorship in the *Federalist papers* (Oakes, 2008; Dale et al. 2000; Mosteller & Wallace, 1964, 1984). The procedure in this method is a combination of the prior probability estimation of some phenomenon (e.g. historical, scientific, or any knowledge from any other field) and the conditional probabilities (new evidence) obtained from the attribution method to make inferential hypothesis about a given disputed text. That is, to determine the authorship of a given text T, if the prior hypothesis or prior probability estimate (say, for example, that there is a 1:3 chance that X wrote T) is approved by the statistical measurement (say, for example, "on", "the", and "up" belonged to X's writing style), the result would be neutral or would support the historical evidence. If the prior hypothesis is contradicted by the statistical measurement, the result would be insignificant (Van Steen, 2012; Forsyth, 2007; Fung, 2000; McEnery & Oakes, 2000; Mosteller & Wallace, 1964).

ii. Cumulative sum charts:

The underlying assumption of cumulative sum charts, also known as cusum charts

(Holmes, 1998; Farrington, 1996; Bissell, 1995; Hilton & Holmes, 1993; Morton & Michaelson, 1990; Morton, 1978; Bee, 1970), is that each individual author has unique distinctive writing habits in writing sentences, which appeared consistently in such features as the usage of nouns, the use of short words (i.e. two or three-letter words), words beginning with a vowel, and a combination of short and vowel words. Distinctive variations in writing habits among different sentences can be taken to be a result of different author(s). Morton (1978) assumed that the rate of occurrence of a writing behaviour for each individual author is consistent and that any significant variation in the proportion of occurrences of the behaviour within a sample of sentences is “prima facie” evidence that the sentences are the utterances of more than one person (Oakes, 2008; Holmes, 1998; Holmes & Tweedie, 1995).

The method requires generation and comparison of two cusum charts, one for the sentence lengths and one for the number of times the stylistic feature or “habit” in question occurs in each sentence. Cusum first measures a particular stylistic feature (say, the number of two-letter words) per a sentence in texts of known authorship and disputed texts and then plots the resulting values, with the vertical axis representing the cumulative sum of deviations and the horizontal the number of sentences. The values for the lengths of the words in n sentences are placed over a curve, each of which is in the form of mean value. This mean value is known as the cumulative sum plot. In simple terms, if we have, say three values for a particular stylistic feature to plot (40.96, 27.42) then we would need to plot them against the cumulative mean of 11.6, 6, and -2.9 (on the vertical scale). This would result in a graph for that stylistic feature showing the deviation of individual values from the mean of that value to that point which is supposed to be distinctive for an individual author; if the cumulative sum plot has a sharp divergence at the point where the texts are joined, then this suggests the authors differ.

However, this method was critiqued and disputed by many researchers of authorship attribution for relying too much on subjective interpretation of the resulting graph and therefore rendering it unreliable for distinguishing between authors. Hilton & Holmes (1993) developed another model based on this method known as “weighted cusums” (see also, e.g., Juola, 2008; Somers & Tweedie, 2003; Somers, 1998; Bissell, 1995) which reduced the subjectivity of interpretation but was found to be still not very accurate compared to other measures. (Stamatatos, 2009; Juola, 2008; McEnery & Oakes, 2000; Holmes, 1998; Holmes & Tweedie, 1995; Sanford et al., 1994; de Haan & Schils, 1993;

Hardcastle, 1993, 1997; Hilton & Holmes, 1993; Canter, 1992).

Today, univariate methods are far less popular in the domain of authorship attribution than they once were. Their limitation is self-evident and has been noted by numerous authors (e.g. Zhenshi, 2013; Forsyth, 2007; Zheng, et al. 2006; Grieve, 2002, Holmes, 1994, 1998; Krzanowski, 1988; Mardia et al., 1979) except perhaps in very special cases. Authorial style is a combination of more or less numerous characteristics, but univariate analysis permits investigation of only one characteristic at a time, the results for different characteristics are not always or even usually compatible, and the consequence is unclear overall results.

More recently, therefore, multivariate methods have increasingly been used. These 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 n -dimensional space, where n is the number of variables used to describe an author's style. There is a large number of cluster analysis methods and a large literature associated with each. An extensive range of these methods is discussed and covered in (e.g. Webb 2002; Duda et al. 2001; Everitt et al. 2001; Everitt & Dunn 2001; Tabachnik & Fidell 2001; Gore, 2000; Grimm & Yarnold 2000; Tinsley & Brown 2000; Gordon, 1999; Jain et al., 1999; Manning & Schütze, 1999; Grimm & Yarnold 1995; Gordon, 1992; Arabie et al., 1992; Kachigan 1991; Hair et al., 1998), as well as in more specialized accounts such as Jain & Dubes (1988) & Gordon (1987). The application of clustering methods to analysis of text corpora is discussed in detail in, for example, Moisl (2015), Mahlberg (2013), Baayen (2008), Lüdeling, and Kytö (2009), McEnery & Wilson (1996).

Until recently, little work has been done using cluster analytical methods with authorship attribution problems. This is understandable, since the domain of stylometry is still at an early stage of development and we can expect expansion in the use of cluster analytical methods as multivariate tools in the resolution of different authorship problems. Holmes (1991, 1992), however, was one of the first researchers to use hierarchical cluster analysis to examine the *Book of Mormon*. A related method is principal components analysis, as applied, for example, by Burrows (1992) to the “Memories of a Lady of Quality” to examine its attribution to Lady Vane. Hierarchical cluster analysis and principal components analysis methods were also used by Dixon and Mannion (1993) to examine the anonymous essays of Oliver Goldsmith. Ledger (1995) used hierarchical cluster

analysis to examine the Letters of St. Paul, and Holmes and Forsyth (1995) also used hierarchical cluster analysis methods to examine the *Federalist Papers*. A related class of methods was used by Mealand (1995), who applied correspondent analysis on the Gospel of Luke and Greenwood (1995), who applied non-linear mapping and hierarchical cluster analysis on the Gospel of Luke and the Book of Acts. Baayen et al. (1996) used principal components analysis to compare the usefulness of several stylometric features like word-based and syntactic-based features for authorship attribution, and so did Stamatatos et al. (1999) to examine a corpus of texts written by various authors of a weekly newspaper. Finally, Merriam (1996) used principal components analysis as the main analytical method to examine Shakespeare's plays, Marlowe's *Tamburlaine* in addition to the questioned *Edward III*. Other studies that used different cluster analytic methods with authorship attribution are, for example, Jockers et al. (2008, 2010), Argamon (2008), Juola (2006), Burrows (2003), and Hoover (2001, 2003, 2004a, 2004b, etc.).

Equally important, a few experimental studies on the accuracy and effectiveness of cluster analysis methods for authorship attribution have been done. For example, Holmes et al. (2001) examined Stephen Crane's and Joseph Conrad's fictions using hierarchical cluster analysis and principal components analysis. The results showed that cluster analysis is able to distinguish between these two writers and also is able to distinguish Crane's fiction from his "shore journalism and New York City journalism". Moreover, the results showed that these two kinds of journalism are different from his war journalism" (Siemens & Schreibman, 2013). Hoover (2001), who did a study to assess the usefulness of cluster analysis for authorship attribution on the basis of the frequencies of the most common words, reported that cluster analysis is able to group works by the same author and distinguish works by different authors with less than 90% accuracy. In 2003, Burrows conducted a study using hierarchical cluster analysis to examine forty long poems written in different genres and originated in the late seventeenth-century by a number of authors. Based on the results, Burrows reported that cluster analysis methods are proven to be the best performing methods in authorship attribution and concluded that "cluster analysis chosen because it offers rather a harsh test of the questions to be considered and also because the family trees in which the results are displayed speak plainly for themselves" (Schreibman et al. 2004: 326; Hoover, 2002).

In addition, the results from these experimental studies also showed that the application of cluster analytical methods for different authorship attribution problems have been rarely

criticized or disputed, except that each clustering method has a 'signature' in the sense that the map or trees it generates tend to have specific characteristics and empirical studies are rarely conclusive (Everitt et al. 2011; Webb 2002; Jain & Dubes 1988).

To sum up, the results from these studies show that cluster analysis is able to distinguish between different authors and different texts of known authorship and disputed texts: works by the same author can be grouped according to their genre or writing styles and authors can be distinguished from one another: the work *x* of author A can be different from or similar to his/her work *y* or work *z*, and the work of author A can be distinguished from the work of author B or author C or disputed work(s) (D, E, F, etc.). Though many evaluative attempts have been made to decide on the best clustering method by using different methods on the same data set and comparing the results, there is no implication that the one particular method or analysis is in any sense 'better' or 'truer' than the others, nor is there any evidence to suggest that one clustering method achieves better and gives appropriate clustering results. In other words, given a range of possibly-different analyses generated by a range of clustering methods, therefore, there is no obvious criterion for choosing among them (Moisl, 2015; Everitt, et al. 2011; Anderberg, 1973).

For more on the authorship studies that considered cluster analysis methods see, for example, Siemens & Schreibman (2013), Juola (2008, 2006), Luyckx (2004), Hoover (20010, Holmes (1998, 1994, 1992, 1991), Holmes & Forsyth (1995), Burrows (1992), Mealand (1995), Greenwood (1995).

All things considered, the discipline of stylometry is still at an early stage of development and has yet to consolidate. There is at present no consensus on what constitutes the core of the discipline. More specifically:

- Despite a very large number of proposed stylistic criteria, there is little agreement on which are valid, and
- Similarly, there is little agreement on which quantitative analytical methods give the most useful and reliable results, and there is again very little work on formal assessment of their validity.

2.2.3 The methodology used in the present study:

As noted in the preceding section, the stylometric literature on authorship attribution distinguishes various categories of problem: closed-class (e.g. Juola, 2008; Binongo, 2003; Diederich et al., 2003; Fung, 2003; Juola & Baayen, 2003; Holmes et al., 2001; Baayen et al., 2002), one-class (e.g. Juola, 2008; Koppel & Schler, 2004), and profiling (e.g. Juola, 2008; Koppel et al., 2002; Stamatatos, 1991). The present discussion is concerned specifically with authorship verification (e.g. Stamatatos, 2009; Koppel et al., 2009; Luyckx & Daelemans 2008; Juola, 2008; Koppel & Schler, 2004).

The problem addressed here is the open or one-class one: given a disputed text and a corpus of works by that author, the aim is to decide whether he or she wrote the text. More specifically, is Coleridge the author of the 1821 Boosey translation of Goethe's *Faust*?

The discussion approaches the problem not by proposing and attempting to justify a hypothesis that he was or was not the author, but by testing an existing one: the Burwick and McKusick hypothesis that he was. This section first outlines the theory of hypothesis testing on which the discussion is based, and then describes the hypothesis testing method which it uses.

2.2.3.1 Hypothesis testing:

In philosophical epistemology or the philosophy of science (Popper, 1959, 1963, 1980; 2002; Chalmers, 1982, 1999, 2007; Ladyman, 2002) there are three main explanations for inferring or explaining a new knowledge about the natural world from a given observed phenomenon (i.e. existing knowledge): deductive inference, inductive inference, and abductive inference. The nature of these types of inference and the differences among them emerge from an example taken from the *Stanford Encyclopaedia of Philosophy* (<http://plato.stanford.edu/entries/peirce/#dia>). Assume the existence of an opaque jar full of marbles.

- Deductive inference:

A deductive inference is one that follows necessarily from given premises or, less

formally, from a given fact or facts: if the given fact or facts are true, an inference from those facts using the rules of logic must also be true. Given our example urn:

Premise: All marbles in the urn are red

Observation: This marble is from the urn

Inference: This marble is red

Given that all marbles in the urn are red, and that one has a marble taken from the urn, it is necessarily true that that marble will be red; if the marble is not red, that is, if the inference is untrue, then either the premise or the observation or both must be untrue. The form of the argument, however, is not in question --it is an absolute rule of logic, and, where it is used, it will always derive true inferences from true premises and observations. A deductive inference from true premises and experimental observations using the rules of logic is always valid with respect to the world.

- Inductive inference:

In inductive inference there are no premises. Instead, an inductive inference is a generalization based entirely on experimental observation of the world: given many or some number of experimental observations, an inference is drawn from them. Referring again to the urn, if someone gives me a sequence of marbles which he says are from the urn, and if all the marbles are red, my inference is that all the marbles in the urn are red. Clearly, this inference is not necessarily true. It may be that there are other colours in there as well, and that it just so happened that all the marbles I saw were red. In other words, inductive inferences are not necessarily true in the way that deductive ones are, and are therefore not guaranteed to be valid with respect to the world.

There are no rules of inductive inference in the way that there are for deductive inference. Instead, we have statistics. Statistics is the discipline that uses sample observations of the world to make inferences about the state of the world, and to assess the probability that such inferences are true.

- Abductive inference:

Like deductive inference, abductive inference starts with premises and makes observations, but the inferences do not result from application of the rules of logic. Going back to the urn, an abductive inference would go like this:

Premise: All marbles in the urn are red

Observation: I have a number of red marbles

Inference: The marbles I have came from the urn

Clearly, the inference does not follow from the premise and the observation, that is, the inference is not necessarily true and therefore not necessarily valid. The inference is reasonable given the premise and the observation, but others are possible --for example, that the marbles I have came from another urn with red marbles in it, or from my pocket.

Deductive inference is the only one of the above three types that guarantees the truth of inferences from existing knowledge. Deductive systems of knowledge exist and are hugely influential in the world, one of them being theology and another mathematics. Such systems are characterized by axiomatisation: certain statements, or axioms, are assumed to be unquestionably true, and all further truths are derived from them via logic, thereby guaranteeing truth preservation relative to the axioms. Unlike theology and mathematics, however, science does not state axioms, and it is not therefore an axiomatic system but instead depends on inductive and abductive inference. It follows that the statements of science are not guaranteed to be true.

The realization that science is not and cannot be a body of truths is fundamental to the currently dominant view of the nature of science, the hypothetico-deductive one associated in the philosophy of science with Karl Popper (1959, 1963), in which scientific research is conducted in a sequence of steps:

1. Some aspect of the real world, that is, a domain of inquiry is selected for the purpose of study.
2. A research question that will substantially further scientific knowledge of the domain is suggested. Given a domain of inquiry, what was the objective of the study the results of which are about to be described? What question did the researcher ask himself or herself?

3. A hypothesis that answers the research question is articulated. Is the answer a new hypothesis? Support for an existing one? Rejection or emendation of an existing one?
4. The validity of the hypothesis is tested by observation of the domain. If the observation is inconsistent with observation the hypothesis must either be adjusted or altered to make it consistent or, if this is not possible, must be rejected. If it is consistent then the hypothesis is said to be held but not proven; no scientific hypothesis is ever proven because it is always open to falsification by new evidence from observation of the domain. With this in mind, 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

On this paradigm, the science of the selected and observed aspect of the domain of inquiry at any given time is a combination of hypotheses that are valid with respect to observations of the domain generated to that time, or, in other words, a combination of best conjectures about what that interesting aspect of the real world is like.

Relative to what has just been said, the proposal that Coleridge was the author of the 1821 Boosey translation is a hypothesis, and there is no hope in principle, much less in practice, of being able to prove the hypothesis true. It is, however, possible to falsify it. If it is falsified then the hypothesis must be abandoned or suitably modified, and if not it is supported but not proven. This discussion attempts to falsify the hypothesis.

A standard approach to falsification is statistical: data relevant to a research question is analyzed using some statistical measure, a hypothesis is framed on the basis of the result, and that hypothesis is tested using the following general procedure (Bhattacharjee, 2012; Gauch, 2003, 2012; Lehmann and Romano, 2005; Lehmann, 1997; Platt, 1964):

1. H_1 is the hypothesis to be tested.
2. An alternative hypothesis H_0 , the null hypothesis, is articulated, which says that H_1 is false.
3. The data is tested using one of a variety of available methods to see if it provides sufficient reason to reject the null hypothesis. If sufficient reason is not forthcoming, then the null hypothesis is accepted and H_1 is rejected. Alternatively, the null hypothesis is rejected and H_1 stands. The latter outcome does not prove the truth of

H1, as one must expect from the foregoing discussion, but says only that, based on the given data, there is insufficient reason to reject it.

This discussion does not, however, use statistical hypothesis testing because the analysis of the relevant data is not statistical. The reasoning which led to the decision not to take a statistical approach is as follows. As noted in the foregoing literature review, many researchers now believe that an author's style cannot be captured by a single or even a few descriptive variables, and that simultaneous analysis of numerous variables is required. That position is adopted here. This means that univariate and bivariate statistical methods are insufficient for present purposes, and that, if statistical methods are to be used, a multivariate methodology is required. The main class of multivariate statistical methods is multivariate regression (e.g. Izenman, 2008; Timm, 2002; Jobson, 1999; Allen, 1997; Berry & Feldman, 1985), which investigates the relationship between more or less numerous independent variables and one or more dependent ones. At an early stage of the research reported here, however, it became clear that selection of sets of independent and dependent variables was problematic: which variables should be independent, which dependent, and why should the sets, once selected, have an independent-dependent relationship? There may well be answers to these questions, but the decision was taken to abandon multivariate regression and to use an entirely different class of methods. In principle, after all, falsification requires only that evidence incompatible with a given hypothesis be identified; that evidence does not have to be statistical in the sense of having been derived from regression analysis.

The class of methods used in what follows all depend on finding structure in a high-dimensional data space, and then using that structure either to formulate or, in the present case, to attempt to falsify a hypothesis. This class includes, among others, principle component analysis, multidimensional scaling, and cluster analysis, and is in the literature sometimes described as statistical. This is a matter of definition. These methods are at best part of descriptive linguistics, and have neither an inferential aspect in the sense of using their results to generalize to a population, nor a set of associated significance tests for the results they generate. For present purposes they are therefore regarded simply as mathematical to avoid confusion with statistics and statistical expectations.

The remainder of this chapter outlines some fundamental concepts relevant to the methods used together with the methods themselves, and specifies the application to

hypothesis testing. They are here characterized as vector space methods for reasons that will emerge.

2.2.3.2 Vector space methods:

This section first presents the concepts of vector space geometry relevant to the discussion, and then shows how they apply to authorship verification.

a. Concepts in vector space geometry:

i. Vector:

In the vector space approach (e.g. Moisl, 2009, 2015; Marshall, 2009; Juola, 2008; Baker, 2005; Rencher, 2012; Singhal, 2001; Belew, 2000; Pyle, 1999; Salton & McGill 1983; Salton et al., 1975) a vector is a set or sequence of n numbers which, when represented horizontally is known as a row vector and vertically as a column vector. Figure (2.1) shows $n = 8$ real-valued numbers, with numerical subscripts denoting each number's place in the sequence: $V_1 = 3$, $V_5 = 8$, and so on. The number n of elements in the sequence is the dimensionality of the vector.

$$V = \begin{array}{|c|c|c|c|c|c|c|c|} \hline 3 & 5 & 6 & 6 & 8 & 4 & 2 & 5 \\ \hline \end{array}$$

1 2 3 4 5 6 7 8

(2.1) An example of a vector

A matrix is a list of vectors. Figure (2.2) shows a matrix M consisting of three 6-dimensional vectors.

		Variables				
		1	2	3	4	5
Cases (data items)	1	3	5	6	8	8
	2	10	6	3	9	2
	3	1	6	4	7	3
	4	9	2	2	5	8

Figure (2.2) data items and variables in a data matrix $m \times n$

The mathematics of vectors and matrices are the foundation of linear algebra, for which see (Marshall, 2009; Datta, 2004; Marcus & Mince, 1988).

ii. Vector space:

In everyday life, we often use the concept of ‘space’ to mean “the boundless three-dimensional extents in which objects and events have relative position and direction” (*Encyclopaedia Britannica*). We understand that we live in a 3-dimensional space within which physical objects have size, shape, and texture and occupy positions and directions. The distances along those positions and directions can be measured and the relative distances between and among objects in the space can be defined and compared with one another. The size or shape or texture of those objects in the space themselves can also be measured and described.

For thousands years geometric attempts were made by early peoples to introduce the notions of space, direction, distance, size, and shape into scientific understanding of physical space or reality, and on the other to solution of practical problems such as construction and navigation. The geometries of ancient Babylonia (2000 BCE-500 BCE) and Egypt (3000 BCE- 500 BCE) passed into the hands of the Greek mathematicians who developed and added to it. From around the sixth century BCE onwards, there were many Greek mathematicians and geometers, among them Thales (635-543 BCE) and Pythagoras (582-496 BCE), and their work culminated in the *Elements of Geometry*

attributed to Euclid (325-265 BCE), which remained the standard textbook on the subject until the 19th century CE (Moisl, 2015; Tabak, 2004).

Developments in mathematics and geometry from the seventeenth-century onwards led to questioning of the fundamental principles of Euclidean geometry both intrinsically and as a description of physical reality, leading to a clear distinction between physical and geometrical space. It was realized that the Euclidean was not the only possible geometry, and alternative ones in which, for example, there are no parallel lines and the angles inside a triangle always sum to less than 180 degrees, were proposed. Einstein used such a non-Euclidean geometry as a more accurate description of curved space-time than was possible with Euclidean geometry. These alternative geometries have continued to be developed without reference to their utility as descriptions of physical (space) reality, and as part of this development the concept of ‘space’ has come to have an entirely abstract meaning which has nothing obvious to do with the one rooted in our intuitions about physical (space) reality. A concept of space under this interpretation is a mathematical set on which one or more mathematical structures are defined, and is thus a mathematical object rather than a humanly-perceived physical phenomenon (Moisl, 2015; Lee, 2010). There are various possible types of space, but the present discussion uses the Euclidean one familiar from elementary mathematics, in which the axes are straight lines orthogonal to one another.

A Euclidean vector space is a geometrical interpretation of a vector in which the dimensionality n of the vector defines an n -dimensional space, the sequence of numerical values comprising the vector specifies coordinates in the space, and the vector itself is a point at the specified Cartesian coordinates (Moisl, 2011, 2009; Baker, 2005; Rencher, 2002; Singhal, 2001). For example, a vector $\mathbf{v} = (2, 4)$ defines a two-dimensional space and its two components are coordinates in that space; a vector $\mathbf{v} = (2,4,6)$ defines a 3-dimensional space, and its values in the specified coordinate system place it at the corresponding position in the space; and so on to any dimensionality. This is shown graphically in Figure (2.3):

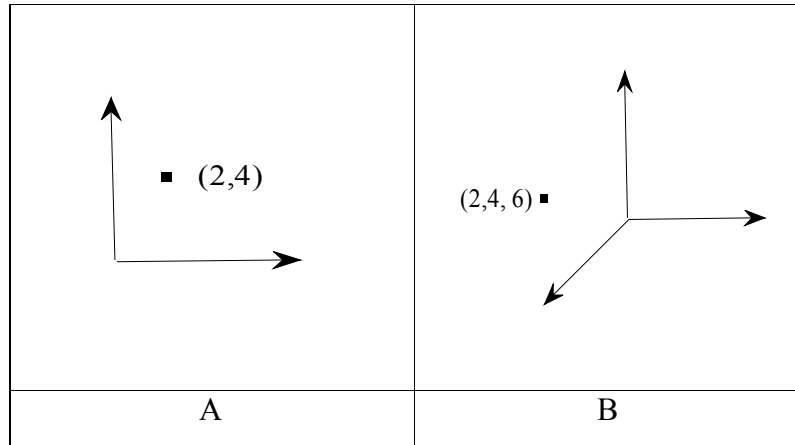


Figure (2.3) 2 and 3-dimensional vector spaces

Any number m of vectors can exist in an n -dimensional vector space, where m corresponds to the number of rows in any given matrix M , and n corresponds to the number of columns.

iii. Proximity in vector space:

In what follows, the generic term “proximity” is used to refer to the distance relations between and among pairs of vectors. According to Moisl (2015, 2011) and Hausner (1965), this may be understood in the following ways.

To speak of a vector as a straight line, we see that if we draw a straight line from the origin $(0,0)$ to the position of any point in the space of the axes (X,Y) , the distance between the origin to that point is known as the length of a vector and can be measured as in Figure (2.4).

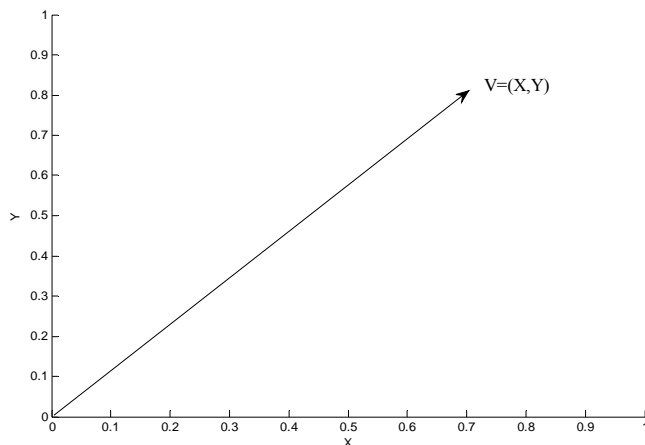


Figure (2.4) A Vector in space

If we draw two straight lines from the origin (0,0) to the position of point A and B then we know that there are two vectors in the space and their lengths can be measured and compared. Two straight lines (vectors) are called equivalent (equal) if they have the same length, and unequal if they have different length. Thus the figure (2.5) shows that the length of vector A is greater than the length of B.

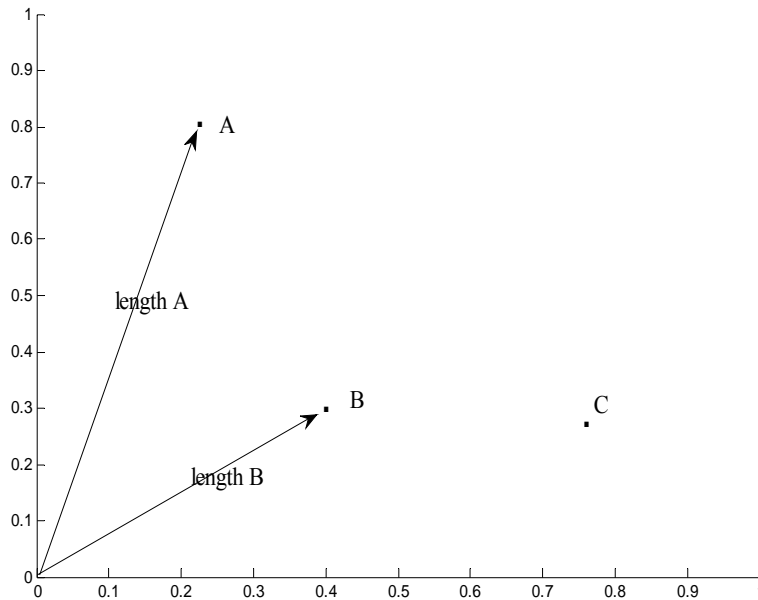


Figure (2.5) Vector length

Because each vector is understood as a straight line determined by 2 points in the coordinate system, we may find the position of any vector if its coordinates are known (i.e. the position of vectors with reference to those two lines is known when we know their distances from the axes). Thus, in the figure (2.5) the position of the vector A is (0.2, 0.8) and vector (B) is (0.4,0.3).

Based on geometrical notions, we may state that the basic elements of vector space are length and angle. These can be used to determine the distance relations between and among vectors, and thus their cluster structure. To illustrate this, when two straight lines (or vectors) meet at a point in a space, there is an angle θ between them, as shown in the Figure (2.6) below.

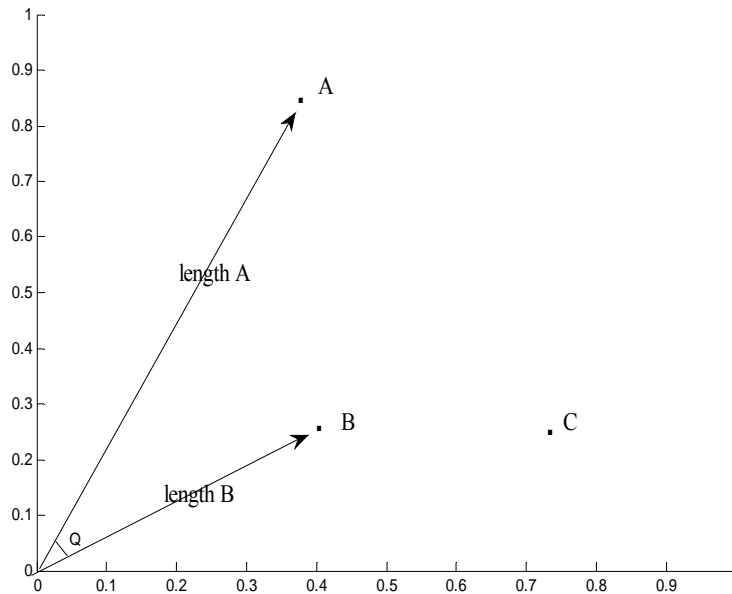


Figure (2.6) the angle between vectors

After the length and angle are identified, the distance between two vectors can be measured and relative distances between pairs of vectors compared, so that distance (AC) in figure (2.7) is greater than distance (AB); this is the basis for several types of clustering methods.

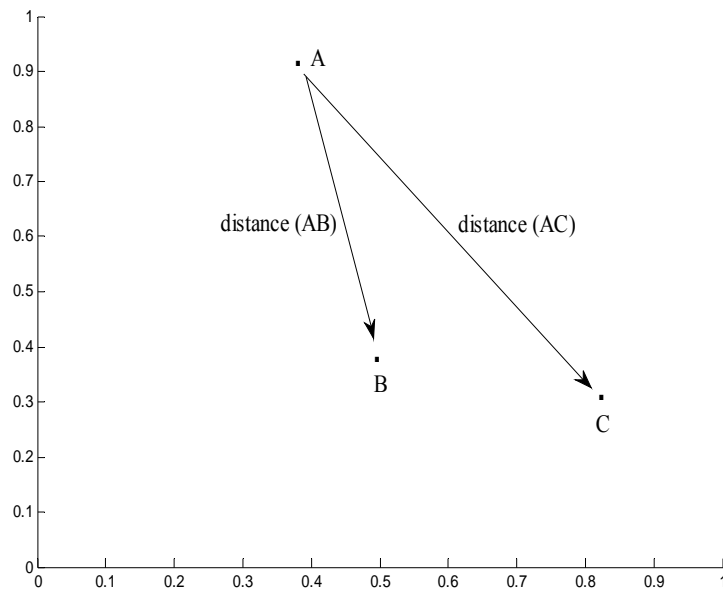


Figure (2.7) Vector distances

The distance between any two vectors in a space is determined by the size of the angle between the straight lines meeting at the main point or origin of the space's coordinate

system, and on the lengths of those lines. Suppose A and B to be any two vectors having identical lengths and separated by an angle θ (figure 2.8):

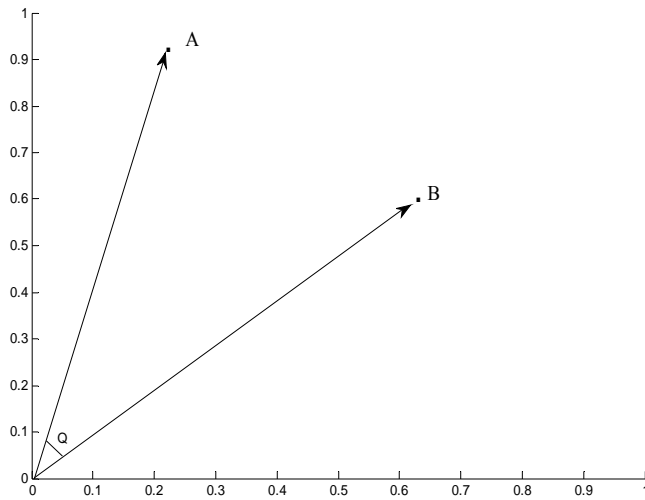


Figure (2.8)

If the angle is fixed and the lengths of the vectors are not the same, then the distance between the two vectors A and B increases (figures 2.9A and 2.9B).

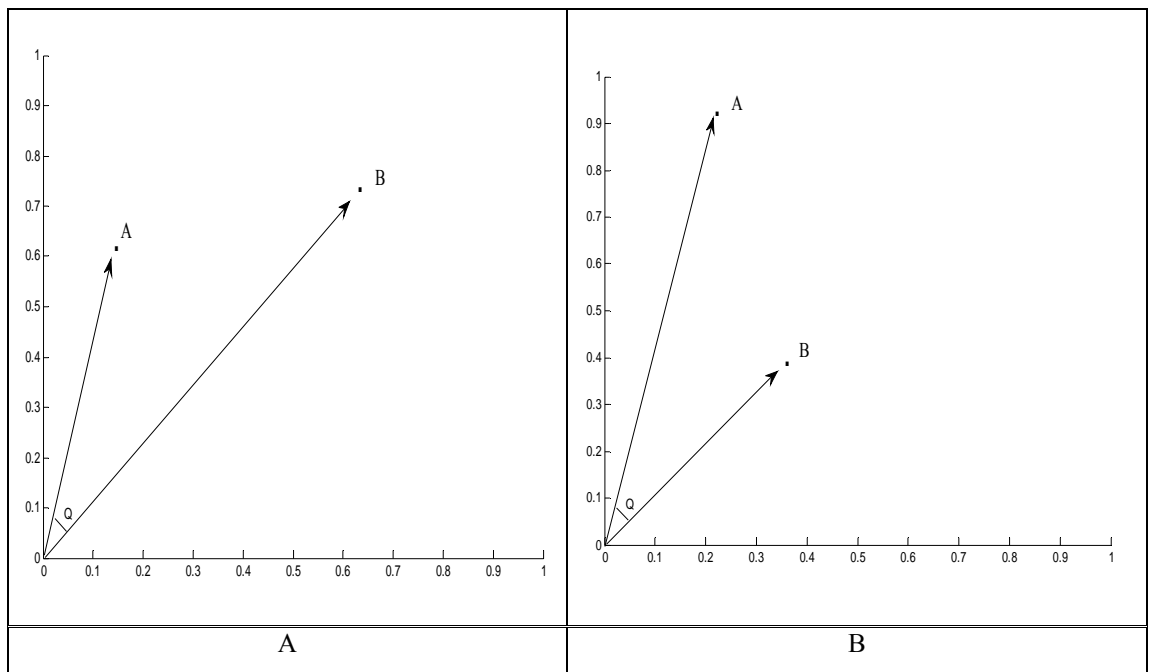


Figure (2.9)

If the lengths of the vectors are the same but the degree of the angle is increased, the distance between the vectors increases (figure 2.10A), and if the degree of the angle is decreased, the distance is also decreased (figure 2.10B).

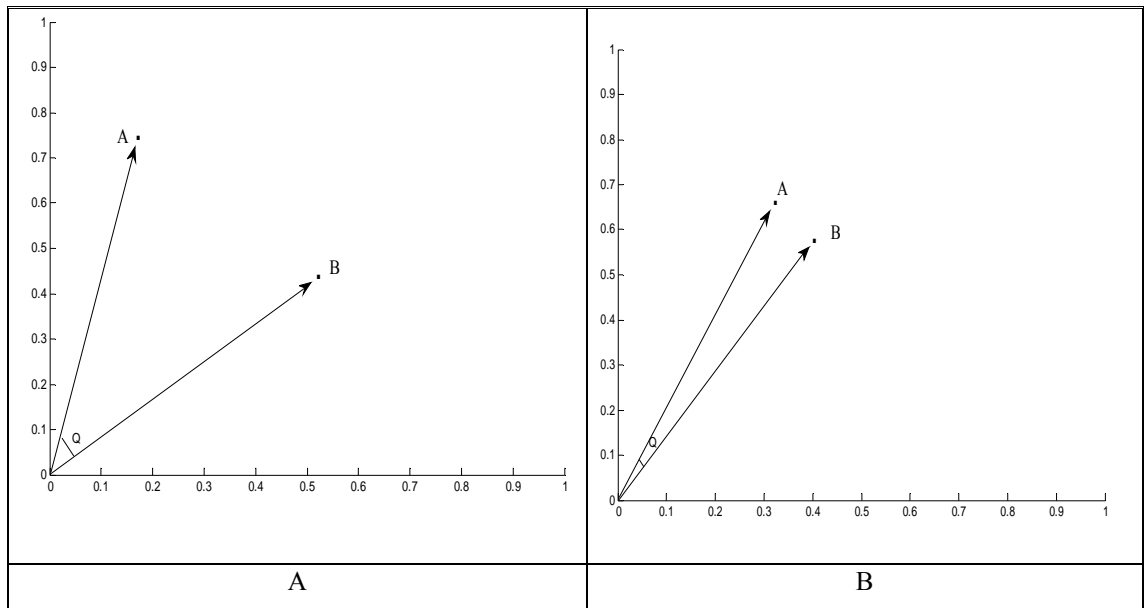


Figure (2.10)

Where there are more than two vectors in n -dimensional vector space, the proximity of one vector to another can be found either by measuring the angle between them or by measuring the distance between them (Moisl, 2015; Everitt et al., 2011, 2001). Angular distance or separation and cosine similarity are examples of the measurement of vectors in a vector space by angle (Moisl, 2015; Everitt et al., 2011, 2001; Singhal, 2001). However, methods of measuring proximity between vectors in terms of the angle between them or distance in Euclidean space are closely related, and if all the variables are measured on the same scale or have been standardized, there is no particular reason to prefer one over another. The measurement of vectors in a vector space by distance is the most common metric measure and is best provided for in software implementations, and so is used here (Everitt et al., 2011, 2001; Hair et al., 2010; Fomby, 2008). Detailed discussion on distances in vector space can be found in (e.g. Deza & Deza, 2009; Xu & Wunsch, 2009; Gan, Ma, and Wu, 2007; and Jain, Murty, and Flynn, 1999).

A note on distance measure is appropriate at this stage. The proximity of vectors to one another is represented in vector space as distance, and such distance can be measured linearly or nonlinearly. In the literature (e.g. Everitt et al. 2011, Jajuga et al. 2003, Gower & Legendre 1986, Gower 1985), numerous distance metric measurements have been proposed which have particular characteristics and can be used in certain applications to calculate the distance from one point to another. These can be divided into two purposely-made types (Moisl, 2015):

1. Linear metrics, where the distance between two points in a space is taken to be the length of the straight line joining the points, or some approximation to it, e.g.: (Squared) Euclidean distance, City block (Manhattan), Minkowski, Mahalanobi, Canberra, Pearson Correlation) and these are available in (e.g. Moisl, 2015, Mooi and Sarstedt, 2011, Everitt et al. 2011, 2001, Deza & Deza, 2009, Fomby, 2008, Lee & Verleysen, 2007, Duda et al. 2001; Gordon 1999; Jain et al. 1999, Lance & Williams, 1966, 1967).
2. Nonlinear metrics, where the distance between the two points is the length of the shortest line joining them along the surface of the manifold and where this line can but need not be straight, e.g. geodesic distance, which mathematically is a generalization of linear in a space (Moisl, 2015, Lee & Verleysen 2007, Gross & Yellen, 2006).

An intuition for how the measure of the distance between vectors in a vector space is best gained by working through a simple numerical example. For present purposes, the distance measure that is most commonly used, most straightforward to apply, and practically simple to understand, will be sufficient. This is the Euclidean distance (Cross, 2013, Everitt et al. 2011, 2001), or straight-line distance, and almost everyone is familiar with, i.e. can be measured with a ruler.

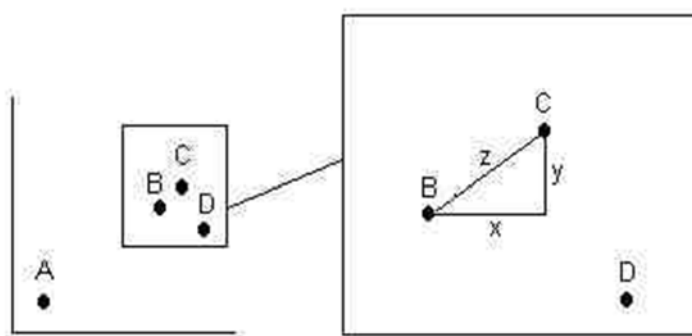


Figure (2.11) Euclidean distance measure

Here the distance between the two points at the vertices of the triangle is the square root of the sum of the squared differences in values for each variable, and mathematically is:

$$\text{Length (z)} = \text{square root } (\text{length}(x)^2 + \text{length}(y)^2)$$

This equation can be extended to include any n vectors in n- dimensional vector space:

$$\text{Length } V_1 \dots V_i = \sqrt{(P_1-Q_1)^2 + (P_2-Q_2)^2 + (P_3-Q_3)^2 + \dots + (P_i-Q_i)^2 + \dots + (P_n-Q_n)^2}$$

Thus:

In two dimensions, if $V1 = (P1, P2)$ and $V2 = (Q1, Q2)$, then the distance is obtained by:

$$\text{Length } V1, V2 = \sqrt{(P1-Q1)^2 + (P2-Q2)^2}$$

In three dimensions, if $V1 = (P1, P2, P3)$, $V2 = (Q1, Q2, Q3)$, and $V3 = (P3, Q3)$, then the distance is:

$$\text{Length } V1, V2, V3 = \sqrt{(P1-Q1)^2 + (P2-Q2)^2 + (P3-Q3)^2}$$

Let $V1 = (2,1)$ and $V2 = (5,6)$ be the given lengths of the sides of the triangle containing the right angle in 2-dimensional space as in figure (2.12) below:

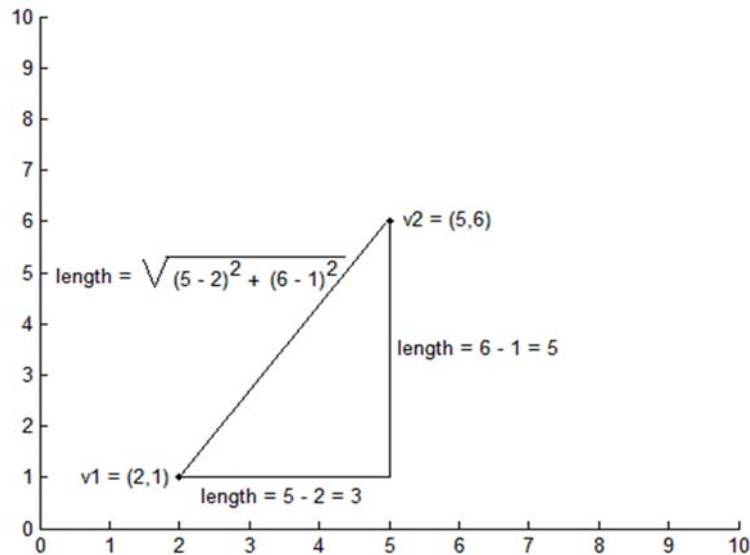


Figure (2.12) Euclidean distance between $v1$ and $v2$

$L(V1, V2) = \sqrt{(5-2)^2 + (6-1)^2} = \sqrt{9+25} = 5.83$ (Euclidean distance), in which this distance corresponds to the length of the line (hypotenuse) which is always opposite the right angle.

However, this distance can be squared to put progressively greater weight on vectors that are further apart and accentuate the degree of separation among them and may help in delineating structures more clearly. (Green et al., 2011)

The equation above can be emended to become squared Euclidean:

$$\text{Length}(z) = \text{length}(x)^2 + \text{length}(y)^2$$

Using this equation and doing the same calculation on the above example, it gives:

$$L(V1, V2) = (5-2)^2 + (6-1)^2 = 9+25=34$$

b. Application to authorship attribution:

In stylometric and as well as in traditional non-quantitative authorship attribution, the concept of similarity between and among texts has been and continues to be fundamentally important (e.g. Moisl, 2015; Oakes, 2014; Koppel, et al., 2002; 2012; Everitt et al., 2011; Lambers & Veenman, 2009; Juola, 2008; Belew, 2000; Hair et al., 1998) where similarity is measured on one or more stylistic criteria such as mean word length or sentence length or vocabulary richness. In vector space terms, similarity is defined in terms of relative distance among texts in vector space. Specifically, given a set $A = \{a_1, a_2 \dots a_m\}$ of m texts by an author A and an anonymous text T :

- A set of n variables to describe the style of A and of T is selected. This defines an n -dimensional vector space.
- T and each of the texts in A are measured in terms of those n variables and the results are stored in a matrix M with $(m+1)$ rows and n columns, such that the value at $M_{i,j}$ (for $i = 1..(m+1)$, $j = 1..n$) is the measurement of text i with respect to variable j .
- The values in each matrix row vector M_i (for $i = 1..(m+1)$) are the coordinates of a point in the vector space, and that point represents text i in the space.
- The relative similarity of any text i to any other is the distance between them.

The distances between the point representing T and the points representing the texts known to be by author A can then be interpreted as measures of stylistic similarity: if T is relatively far from those of A then the hypothesis that A is the author of T is falsified relative to the stylistic criteria used, and confirmed, though not of course proven, if relatively close.

The twin ideas of using relative distance in vector space as a measure of textual similarity and of using this relative distance as a criterion for authorship attribution is fundamental to the methodology of the present discussion. The remainder of this chapter gives a

detailed exposition of how this methodology is implemented.

2.2.3.3 Data creation:

a. Variable selection:

For vector space based hypothesis testing to be effective, the set of n variables used to describe the relevant texts is required. Intuitively, such a variable is a valid descriptor which captures a salient aspect of textual style, given some definition of salience; in vector space terms, the set of descriptors must consistently locate texts known to be similar near to one another in the n -dimensional space, and texts which are known to be dissimilar far apart.

Unfortunately, stylometry cannot at present offer an agreed-upon definition of stylistic salience, and this has generated the numerous stylistic descriptors surveyed above. The problem has been that, while most of the proposed descriptors are intuitively plausible, in most cases their effectiveness has not been assessed relative to objective criteria such as, for example, their reliability in distinguishing texts by different authors from one another. Using such unassessed descriptors renders the significance of analytical results imponderable.

More recent work has attempted to identify and assess reliable stylistic descriptors (e.g. Oakes, 2014; Ramezani, et al., 2013; Luyckx & Daelemans, 2008; Abbasi & Chen, 2008; Juola, 2008; Stamatatos, 2008, 2009; Cyran & Stanczyk, 2007; Grieve, 2007; Feiguina, & Hirst, 2007; Bozkurt, et al., 2007; Miranda & Martin, 2007; Forsyth, 2007; Zheng et al., 2006; Koppel et al., 2006; Argamon, et al., 2005; Luyckx, 2004; Koppel & Schler, 2003; Binongo, 2003; Burrows, 2002a; Love, 2002; McEnery & Oakes; 2000; Forsyth & Holmes, 1996; Holmes, 1998). In particular, to determine which stylistic descriptors are best relative to authorship attribution, Grieve (2007) assessed the effectiveness of 39 different stylistic descriptors in a large collection of 40 texts matched for genre (the Telegraph newspaper opinion column) and time period (2000-2005) written by authors from similar social backgrounds (middle-aged, conservative, Anglo-Saxon, upper-to-middle class, well-educated, British) for the same audience (the readership of the Telegraph's opinion section). Grieve found that function words are the most effective stylistic descriptors to attribute authorship to disputed texts. This finding is also consistent

with the existing conclusions from previous authorship experiments performed to evaluate and assess various stylometric features (e.g. Ramezani, et al, 2013, Bozkurt et al., 2007; Argamon & Levitan 2005; Baayen et al., 1996).

The explanation of the effectiveness of this criterion is typically that function words are on the one hand independent of textual content, and on the other are indirect syntactic markers (e.g. Kestemont, 2014; Zhenshi, 2013; Yu, 2012; Smith, 2008; Chung & Pennebaker, 2007; Koppel et al., 2007; Koppel et al., 2006; Merriam, 2006; Riba & Ginebra, 2005; Girón et al. 2005; Zhao & Zobel, 2005; Hoover, 2001, 2004; Binongo, 2003; Burrows, 1987, 2003; Yang, 1999; Holmes, 1992, 1994, 1998; Wallace, 1984, 1964). This seems plausible.

The fundamental hypothesis underlying stylometry in general and authorship attribution in particular is that an author's style can be characterized by his or her lexical selection preferences and the arrangement of selected words into syntactic structures, and that style so defined varies between and among different authors. This can be understood as:

- Content vs function words:

Content words such as nouns, verbs, adjectives, and adverbs have semantics related to particular topic domains and situations, i.e., farming, computer science, etc. The semantics of function words, on the other hand, are independent of topic domains and situations. If the aim is to classify documents by topic, then content words would be used. If the aim is to classify documents by author independently of topic, then content words should not be used because they confuse the issue. What remains is function words.

- Function words as syntactic markers:

Ideally, any stylometric analysis would include varieties of syntactic usage as criteria. Where parsed corpora are unavailable, however, function words often mark syntactic usage indirectly. There are distinct categories of function words for grammatical use and their presence indicates particular constructions (Smith and Jong, 2005). For example, use of relativizers as indicator of dependent clauses and thus of degree of

syntactic complexity, prepositional phrases as opposed to possessives ('the road's end' / 'the end of the road') etc.

Today's literature on the use of function words approach for examining the authorship of disputed texts is abundant. In it, function words approach is addressed as one of the most useful and suitable stylometric criteria when the purpose is to capture styles of writing or what is known as authorial characteristics of an author. A number of explanations exist in support of this based on different research outcomes and assumptions. Some of these are:

1. Research in words use and writing styles (e.g. Montague, 2011; Chung & Pennebaker, 2007; Smith & Jong, 2005; Kennedy, 2003; Baayen et al., 1995; Smith & Witten, 1993; Caplan, 1987) shows that the average persons' everyday vocabulary consists of about 10.000 words. It also shows that there are about 250-400 function words in English, each of which has approximately 20 distinct uses or may be more. The twenty most common words (the, and, of, a, in, to, it, is, was, that, this, have, with, for, not, on, as, do, you, I) alone make up almost 25 percent of all the words we use every day (Chung & Pennebaker, 2007; Kennedy, 2003; Zipf, 1965). As a result of this, it is thus reasonable to predict that function words are used across all topic domains, situations, and styles of writing. It is also reasonable to predict that all authors are bound to use function words in all writing situations and contexts and that all are expected to leave distinctive function word usage traces on text written by them, which stylometrists try to capture to distinguish one author, or text, from another.
2. Function words are resistant to stylistic imitation and forgery. This is based on the assumption that function words are (or assumed to be) outside the conscious control of an author (e.g. Kestemon, 2014; Stamatatos, 2009; Koppel et al., 2009; Juola, 2008; Chung & Pennebaker, 2007; Argamon & Levitan, 2005; Peng et al., 2003; Binongo, 2003; Holmes, 1985, 1994; Garrett, 1982). The way our brains work to use function words during sentence formation in fact differ from one person to another, and this makes it difficult to memorize function words usage of others or even ourselves. A great deal of research supports this assumption (e.g. Fromkin, et al. 2014; Pennebaker, 2011; Fernández & Cairns, 2010; Chung & Pennebaker, 2007; Crane, 2001; Lancashire, 1997, 1998; Meyer, 1979; Bailey, 1979) on the basis of the fact that there is no definite proof that our brain is equipped with having control or memory to imitate other's or someone else's stylistic use of function words.

3. Function words are part of an author's style. Two explanations are given for this assumption (e.g. Stamatatos, 2009; Koppel et al., 2009; Juola, 2008; Grieve, 2002; Holmes, 1994). First, the way an author uses or selects a set of function words is determined by the presence of certain stylistic patterns or internal structures in a text at hand, which allow him/her to select between the variant structures and the semantically equivalent variants. An author may have a preference for certain syntactic constructions, say, ("to + verb" or "passive voice") which requires certain set of function words, say ("to" or "by" or "is") and, at the same time, this preference may also depend on the meaning that this set of function words conveys. If he/she replaces one function word with another there will be a sentence with different meaning. This is essentially the reason that Mosteller & Wallace gave (Argamon et al. 2005; Grieve, 2002) when wrote "we need variables that depend on authors and nothing else...some function words come close to the ideal" (Mosteller & Wallace, 1984:266). The other, though they are not entirely without meaning, function words are assumed to be topic-independent in the sense that a set of function words which an author uses to express structural relationship with other words in a sentence should be the same regardless of whether he/she is describing religious sermons or political speeches (e.g. Guerra, et al., 2013; Pennebaker, 2011; Chung & Pennebaker, 2007; Kestemont, 2000; Smith & Written, 1993; Damerau, 1975). This could also be one of the reasons why function words are mainly used in preference to content words in authorship studies because they are not biased by the content or genre of an author's writings (Stamatatos, 2009; Koppel et al., 2009; Argamon and Levitan, 2005; Damerau, 1975; Zipf, 1949).

4. Authors using English language in any period of time tend to use the same function words at stable rates in texts written by different styles and on diverse topics, (Stamatatos, 2009; Koppel et al., 2009; Juola, 2008; Holmes, 1994). The relative frequency of function words tends to be stable within an author's own work and between works by the same author but tends to vary greatly within works written by different authors and within different genres (Chung & Pennebaker, 2007; Argamon and Levitan, 2005). If we accept the idea that function words are essential to the way authors write to connect words, phrases, or entire clauses together and the assumption that function words flow straightaway from an author's mind, then we would expect to see differences in their function word frequencies. But the results of different studies show that the usage's rate of function words in fact varies from author to author. For example, if an individual author habitually tends to use the construction 'to + my +

N./adj.’ across a number of texts, we can expect to have a remarkable stability in the rate of use of ‘to’ or ‘my’ within these texts and also a slightly higher frequency for ‘to’ or ‘my’ than other authors, or if someone else habitually tends to use the phrase ‘as far as’, we can expect to have a remarkable stability in the rate of use of ‘as’ and also a higher than average frequency ‘as’ across a number of texts.

5. Because they only make sense in relation to other words, function words tend to demonstrate very high frequency usage; an advantage that allows stylometrists to quantify a large number of measurements. (Stamatatos, 2009; Culpeper, 2002; Thomas, et al., 2004; Grieve, 2002; Barker, 2000; Kilgarriff & Rosenzweig, 2000; Enkvist, 1964, 1973).

Whatever the explanation, however, the work of Grieve and other cited above indicates that function words are currently the best criteria for discriminating different authors, and for that reason they are used as variables for constructing the data in the present discussion.

b. Matrix construction:

The m texts in any given study are represented by m rows of a data matrix D , with each row representing a different text, and the n function words selected as variables are represented by the n columns of the matrix. What should the matrix values be? In principle, any one of an unbounded range of value types can be used. The value at D_{ij} might, for example, be the standard deviation of variable j in text i , or a binary value where 1 represents the occurrence of variable j in text i and 0 its non-occurrence, or anything else considered to be a useful measure of function word distribution in the texts being studied. In practice, the frequency of textual features of interest is used almost exclusively in the stylometric literature. Why frequency? To judge from the literature, the answer appears to be based on the intuition that writers differ in the frequency with which they use stylistic features: one writer might have a tendency to use long sentences with many dependent clauses and another to avoid subordination, for example, or might be inclined towards frequent use of adjectives and another to a spare style that avoids adjectival description, and so on. Counting such features is held to be an intuitively reasonable way to describe style, and that is the position taken here. Specifically, the value at any location D_{ij} in the matrix analyzed in this study is the number of times

function word j occurs in text i .

c. Data optimization:

Once the data matrix has been constructed, two transformations are required prior to its analysis: normalization and dimensionality reduction.

i. Normalization:

Where there is more than one text in a corpus, as is usual, it might happen that all the texts are equal in length. If they are not, however, a major problem for cluster analysis arises that must be resolved. In what follows, we will look at what this problem is, and what the solutions are.

The essence of the problem is that, where the data matrix is based on variable frequency, the token frequency of any given variable will, in general, increase in at least approximate proportion to document length: the frequency of, say, the function word 'the' will be much higher for a novel than for a short email message. This means that, again in general, shorter texts will have smaller word-frequency occurrences than longer ones, which in turn invalidates any analysis which directly compares the rows of the data matrix representing the varying-length texts (Moisl, 2009a; Baker, 2001; Holmes, 1998; Baayen, 1996; Baayen et al., 1996). To see why, consider an analysis of a corpus whose constituent texts vary in length, the aim of which is to distinguish the texts stylistically on the basis of the frequency of occurrence of the pronoun 'I'. Say there are 50 occurrences of 'I' both in texts A and B. Knowing only these frequencies, one would judge that the two texts A and B are identical on this criterion. But Text A is 50,000 words-long, and text B is only 500. It is clear that, though they both have the same number of occurrences of 'I', the significance of their respective frequencies is far from identical: the personal pronoun 'I' is relatively infrequent in text A in the sense that its probability of occurrence is only $50/50000$, and relatively frequent in text B because its probability of occurrence is $50/500$, or one hundred times as great; if text B had also been 50,000 words long instead of 500, the frequency of 'I' would, on the basis of its observed probability, have been $100 \times 50 = 5000$ occurrences, and on that basis the two texts A and B would be judged as very different on the personal pronoun 'I'.

Variation in document length is, in short, a problem for any analysis which aims to distinguish documents on the basis of the frequency of occurrence of selected descriptor variables. The effect on cluster analysis specifically can be exemplified by means of an example. Assume the existence of a corpus C consisting of m varying-length documents, and a data matrix D abstracted from C in which the m rows represent the documents, the n columns represent whatever variables have been chosen to describe the documents, and the value at D_{ij} is the frequency of occurrence of variable j in document i . D is cluster analyzed using one of the methods, hierarchical analysis, described later in this chapter, and the result is shown in Figure (2.13).

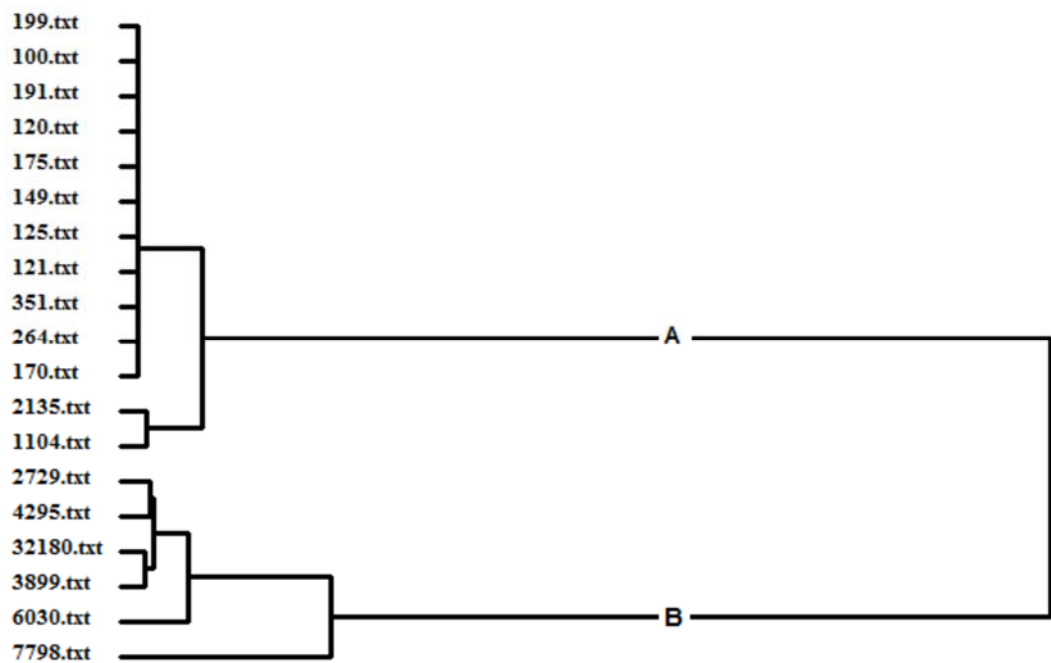


Figure (2.13) text-length based clustering

A hierarchical cluster analysis generates a binary tree structure. How such a tree is constructed is discussed in detail below; for present purposes, it is sufficient to note that the tree represents the structure of similarity relationships among the objects being compared in terms of the variables selected to describe the objects. The relative lengths of the lines joining subtrees represent the relative similarities of the subtrees: the longer the horizontal lines joining and two subtrees, the more dissimilar the texts in the respective subtrees are. Thus, in Figure (2.13), the texts in the subtree labelled A and those in B are relatively very dissimilar to one another, those in C and D are less dissimilar, and so on. The numbers at the leaves of the tree represent the lengths of the texts in C in terms of the total number of words they contain, and, as is readily seen, those texts have been clustered strictly on the basis of length, with the shorter ones in subtree A and the longer ones in B, and the constituents of A and B similarly sub-clustered. Relative document

length has, in other words, obscured any stylistic similarities which the texts may have. This effect applies, moreover, to cluster analysis of frequency matrices abstracted from varying-length document corpora generally (Moisl, 2009b, 2011, 2015).

It is clear that the effect of variation in document length for cluster analysis of frequency matrices must be mitigated or eliminated before the analysis is undertaken. We shall see that the documents cluster analyzed in the present study vary substantially in length, and a method of eliminating this as a factor in the analysis is therefore presented in what follows.

One solution to this problem is simply to adjust the lengths of all the texts in the corpus of interest so that they are identical, either by adding more material to the shorter texts, or by truncating the longer ones, or by a combination of the two. This is obviously unsatisfactory: shortening longer texts loses information, and lengthening shorter ones raises the twin questions of what text should be added, and of the consequent effect on document validity. This solution is not further considered here. The alternative is to adjust the data matrix abstracted from the corpus in such a way as to eliminate variation in document length as a factor affecting the frequencies.

The literature on document length normalization (e.g. Moisl, 2009b, 2008, 2011, 2015; Manning, Raghavan, and Schütze, 2008; Greengrass, 2001; Belew, 2000; Spärck-Jones et al., 2000; Baeza-Yates & Ribeiro-Neto, 1999; Singhal et al., 1996; Singhal, Buckley, & Mitra, 1996; Singhal et al. 1995, 1996a, 1996b; Robertson & Spärck-Jones, 1994) contains a variety of methods, not all of which need to be described in detail here. Instead, normalization by mean document length is used for the present analysis, as developed in (Moisl, 2011), both because of its intuitive simplicity and because it does what is required at least as well as other methods.

Mean document length normalization involves transformation of the row vectors of the data matrix being analyzed in relation to the average length of documents in the corpus from which the matrix was abstracted:

$$M_i = M_i \left(\frac{\mu}{\text{length}(C_i)} \right)$$

Where:

- M_i is the matrix row vector representing the frequency profile of text C_i

- Length C_i is the total number of lexical types in C_i
- μ is the mean number of lexical types across all texts in C , and is obtained by:

$$\mu = \sum_{i=1..m} \frac{\text{length}(C_i)}{m}$$

- The values in each row vector M_i are multiplied by the ratio of the mean number of lexical types per text across the collection C to the number of lexical types in text C_i .

The effect is to decrease the values in the vectors that represent long texts, to increase them in vectors that represent short ones, and, for texts that are near or at the mean, to change the corresponding vectors little or not at all. To exemplify this method, let M be a matrix having 3 documents (a, b, c) with unnormalized values of four lexical types as shown in Table (2.2).

	V1	V2	V3	V4
	the	a	you	oh
doc.a (length= 500)	12	15	3	53
doc.b (length=1500)	4	36	1	36
doc.c (length=2430)	7	80	0	29

Table (2.2) Matrix M of 3 length-varying documents and 4 unnormalized frequency profile

Applying the mean document length normalization formula:

- Find the mean length across all documents. Thus we have $500 + 1500 + 2430 / 3 = 1476$
- In each row vector, the count for a given lexical type is multiplied by the mean document length, then divided by the total number of frequency counts occurring in that row vector. Thus:

For document (a):

$$12 \times (1476/500) = 35.42$$

$$15 \times (1476/500) = 44.28$$

$$3 \times (1476/500) = 8.85$$

$$53 \times (1476/500) = 156.45$$

For document (b):

$$4 \times (1476/1500) = 3.93$$

$$36 \times (1476/1500) = 35.42$$

$$1 \times (1476/1500) = 0.98$$

$$36 \times (1476/1500) = 35.42$$

For document (c):

$$7 \times (1476/2430) = 4.25$$

$$80 \times (1476/2430) = 48.59$$

$$0 \times (1476/2430) = 0$$

$$29 \times (1476/2430) = 17.61$$

Transformed in this way, the resulting matrix looks like Table (2.3):

	V1	V2	V3	V4
	the	a	you	oh
doc.a (length= 500)	35.42	44.28	8.85	156.45
doc.b (length=1500)	3.93	35.42	0.98	35.42
doc.c (length= 2430)	4.25	48.59	0	17.61

Table (2.3) Matrix M of 3 documents length-normalized frequency profile

The effect of normalization is clear: all the values in the document (a) have been substantially increased because it is significantly shorter than the mean document length: length-500 < 1476 (the mean). For document (b), the values have been slightly decreased because it is slightly longer than the average document length: length-1500 > 1476. Finally, the values for document (c) have been substantially decreased because it is

significantly longer than the average document length: $2430 > 1476$.

More on document length normalization can be found in Moisl (2015, 2011, 2009b), Priddy and Keller (2005), and Singhal et al. (1995, 1996).

ii. Dimensionality reduction:

As noted in the forgoing discussion of vector space geometry, stylistic descriptors, or variables, are represented by vectors in n -dimensional vector space and vectors themselves are data points distributed in the space. The dimensionality n of the space is the number of variables. Technically, any data set with dimensionality greater than $n = 3$ is called multidimensional multivariate (Arppe, 2008; Chan, 2006; Bartke, 2005; Rencher, 2002). However, the conceptual boundary between low and high dimensionalities is not always precisely stated, and therefore used in a loose manner: high-dimensional data is usually used to refer to any dimensionality greater than 4, and therefore, a set of data in 2, or 3 or 4 dimensional space can be generally referred to as a low dimensional data. Nevertheless, the researcher technically reserves the term high-dimensional data for dimensionalities greater than 3. More information on this can be found, for example, in Wing & Chan (2006), Bartke (2005), Oliveira & Levkowitz (2003).

The conceptual boundary related to human's visual perception capabilities and intuitive understanding can be an obstacle with respect to higher-dimensional spaces (Zeng et al. 2011; Moisl, 2009a, 2009b; Chan, 2006; Bartke, 2005; Rencher, 2002): humans find it difficult, at the very least, to conceptualize 4 or 5 dimensional spaces, and it seems impossible to do so for, say, a 50-dimensional one. Mathematically, however, there is no problem with spaces of dimensionality greater than 3; this study will deal with such spaces from a mathematical point of view, using the intuitions based on human experience of a 3-dimensional world as a metaphor when required for conceptual clarity. High data dimensionality is a problem for cluster analysis and needs to be reduced as much as possible to enhance the reliability of clustering results. The section first considers the nature of the problem and then describes several ways of resolving it.

- The problem of high dimensionality in data:

It was noted earlier that any number m of n -dimensional vectors can exist in an n -

dimensional vector space. Geometrically, such a collection of m vectors is called a manifold. The manifold, for example, in figure (2.14a) is, or assumed to be, a straight line since there are only two vectors in the space. The manifold in figure (2.14b) is a curved line since there are three vectors embedded in the space. The manifold in figure (2.14c) is a complex shape due to plotting a large number of vectors in the space.

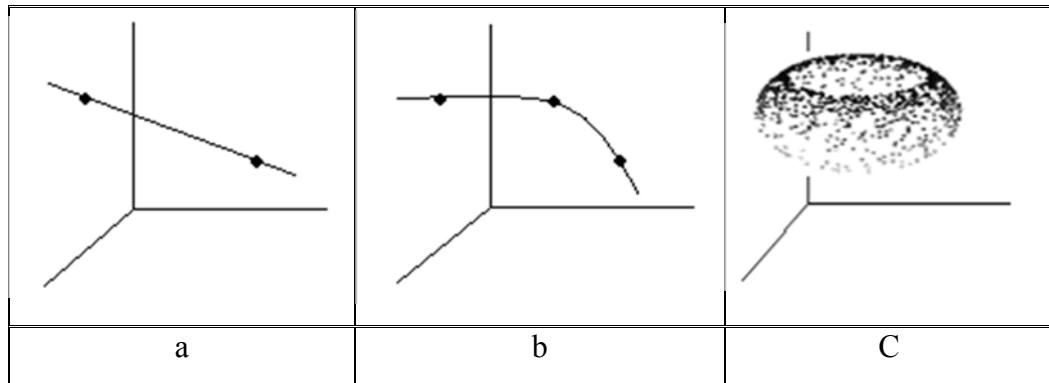


Figure (2.14) Categories of manifold definition

The key observation is that there are always restrictions of some kind on the shape of the manifold even when it may be well-defined by the vectors. Some of these problems (Pyle, 1999: 84) are:

- The vector points are not well-situated or located on some part of a manifold's n -dimensional surface; clustering in that place is not likely to give satisfactory results.
- In another space there may be very few vector points situated in dimensional space to define the shape of the manifold. Here, if we observe the manifold points, the results might be unsatisfactory for a reason different from the one described above.
- At other places the shape of the manifold may be well defined by the vector points, but have complicated shapes. For example, complicated or problematic shapes may qualify manifolds having a hole or tunnels through them or a fold over themselves. Many analytical and projection methods simply fail to deal with such a shape.

By this point in the discussion (i.e. the shape of the manifold), it is often the case that to discern the shape of the manifold there must be enough vectors lying or populating in a manifold embedded in the Euclidean space to enhance or enrich it and, therefore, give it adequate definition (Moisl, 2009a, 2009b). But here is a problem with this claim. The essence of the problem with data manifolds in high-dimensional spaces with respect to

cluster analysis is that, as dimensionality increases, it becomes increasingly difficult to get enough data vectors to define the manifold well enough for cluster analysis to give reliable results. To see why this is so, consider what happens to the size of a cube as dimensionality is increased from 3 to 100, where size is measured in terms both of volume and of the length of the diagonal from the origin to the opposite corner.

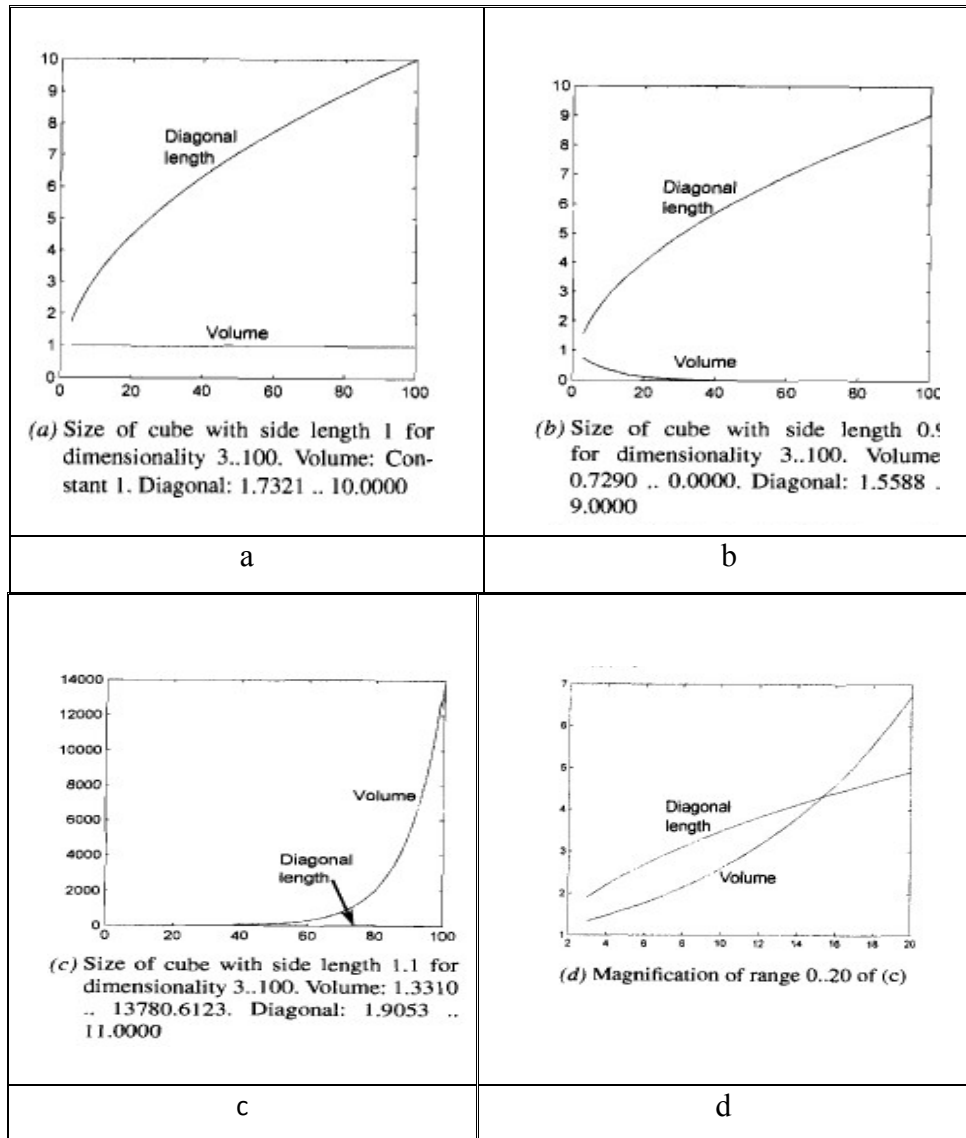


Figure (2.15) Effect of dimensionality increase on the size of a cube

By looking at figures (2.15a-2.15d), which are reproduced from Moisl (2015:73), our observation is that high dimensional spaces show highly counter-effective properties:

- In figure (2.15a), we have diagonal and volume where the length of the diagonal grows even though the volume remains constant.

- In figure (2.15b), we have diagonal and volume where the length of the diagonal grows even though the volume converges to 0.
- In figure (2.15c) and figure (2.15d), we have diagonal and volume where the volume quickly starts to increase at a much greater rate than the length of the diagonal.

The general conclusion therefore is that when dimensionality increases, counter-intuitive influence becomes even more dominant. As a result of this effect, our intuitive anticipation based on experience of the three-dimensions of space suggests that there should be balance or proportion between volume and diagonal length regardless of the scaling of the data values, but that is not the problem: rescaling of the data to axis lengths that are less than, equal to, or larger than 1 fundamentally changes their relationship. Volume is a human intuition based on experience of the higher-dimensional objects, and the mathematical formulation of it identifies the intuition for dimensionality 3. Beyond that dimensionality volume becomes intuitively meaningless, and the mathematical formulation of it reduces to the well-known effects of multiplying values less than, equal to, or greater than 1 n times (Moisl, 2015: 72-3).

Working with high-dimensional data means working with data that are populated in high-dimensional spaces (Verleysen & Francois, 2005). To understand this, some discussion is required.

As a matter of fact, multidimensional multivariate data analysis is all about finding a suitable projection or mapping to represent the data vectors in a visual form (i.e. preferably 2-dimensional space) to find interesting structures of dis/similarities and create hypotheses. While the visualization of data vectors on a 2-dimensional space seem very straightforward and efficient, the visualization of high-dimensional vectors in n -dimensional space becomes harder (Moisl, 2015; Sakai and Hashimoto, 2011; Everitt et al. 2011; Lee & Verleysen, 2007; Bartke, 2005; Belew, 2000). To see why, if trivariate data set, say, of 1000 3-dimensional vectors are plotted in 3-dimensional space, the data points would be literally scattered around the space like a cloud, as in Figure (2.16):

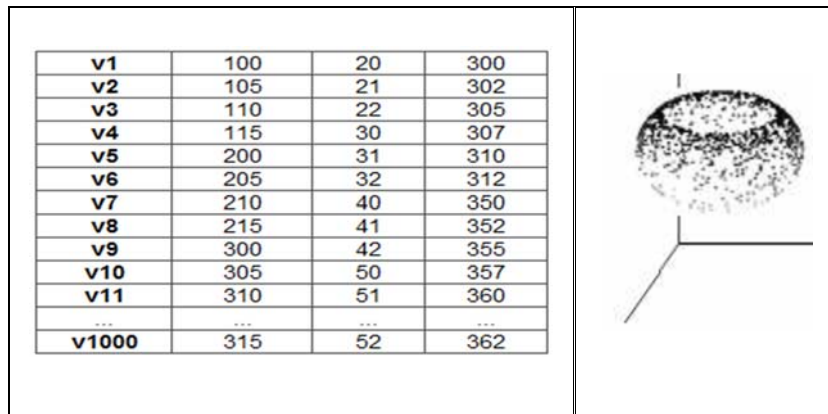


Figure (2.16) Data set of 1000 vectors in 3-dimensional space shown as cloud of data points

With multivariate data set, the visualization becomes even more difficult or uneasy even if they are plotted on 2-dimensional vector space. The data vectors or the distance between them are far too close to each other in the space or are sufficiently unique to visualize or identify any existing patterns, as in Figure (2.17).

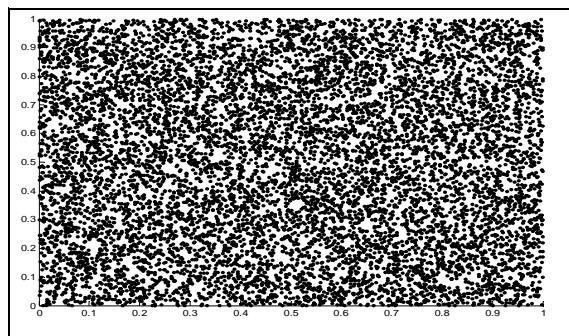


Figure (2.17) Plots of very large vectors in 2-dimensional space, where a pattern is hidden by the number of vectors

More specifically, in dealing with high-dimensional data, we experience difficulty in understanding or conceptualizing:

1. The interrelationships of variables within a single data item: how, for example, are function words ‘he’, ‘by’, ‘you’, ‘now’, and ‘for’, of any given text interrelated?
2. The interrelationships of variables within complete data items: how do texts measured on the above function words compare to one another?

because high-dimensional spaces have a number of geometrical properties which have a large influence on the performances of cluster analysis methods. Some of these properties are as follows:

For a fixed number of data vectors m and a uniform and fixed variable value scale, the manifold becomes increasingly sparse as their dimensionality n grows (Moils, 2015). The question that needs to be addressed here, as one might ask, what does it mean to say that “the manifold becomes increasingly sparse as their dimensionality n grows”? To understand the significance of this, we assume that the larger the data vectors, the better or the clearer the shape of the manifold points. If this is the case, then we need to use either enough data vectors or collect more vectors to adequately fill the empty space and define the shape of the manifold points. To see the problem, the discussion considers these two alternatives in detail. Taking the first alternative, that is, getting enough or a fixed number of data vectors to fill the space is usually difficult or even intractable as its dimensionality grows, however. Since the shape of the manifold points is all about the distribution of vectors in the space, the fundamental problem is that when the dimensionality becomes larger, the volume of the space becomes larger as well but very quickly that the locally embedded vectors in the space becomes sparse, and to improve the manifold definition, more and more vectors are required until, equally quickly, getting enough becomes impossible (Moisl, 2009a, 2009b; Bellman, 1961). To illustrate this and see what happens, suppose that a 2-dimentional space be the given coordinates X and Y in which each dimension or coordinate has 10 cells in the intervals range from 0...9 and consider a collection of 5 two-dimensional vectors such as (1,9), (9,9), (3,5), (6,1), (4,2) to be distributed in it. It is intuitively clear that the cells should cover all the existing vectors and still allow extra 95 ones; hence the vertical and horizontal axes go from 0 to 9, we may thus calculate the whole space as 10^2 to stand for $10 \times 10=100$ locations, that is, there can be a maximum of 100 vectors in this space, as shown in Figure (2.18) below:

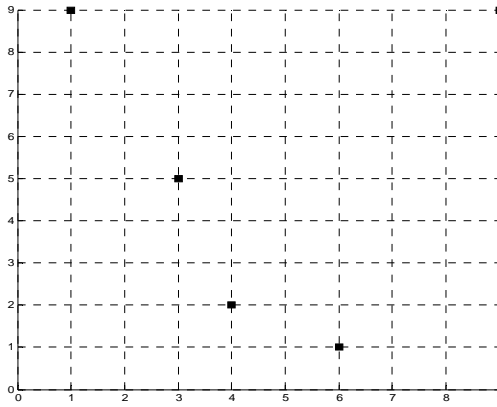


Figure (2.18) Five 2-dimensional vectors in space

Finding the number of spaces for possible vectors may simply be expressed in scientific notation as something like r^d function where r is the measurement or scale range $1 \dots n$ (here $0 \dots 9 = 10$) which is raised to the power of d , the dimensionality. This formula is a very simple way of showing a serious problem of the rapid increase of the vector space size with dimensionality. For a collection of 2 three dimensional vectors such as $(0, 9, 2)$ and $(3, 4, 7)$ with all three axes are in the same range $0 \dots 9$ (i.e. 10^3 or $10 \times 10 \times 10$), there will be 1000 possible vectors, as in Figure (2.16):

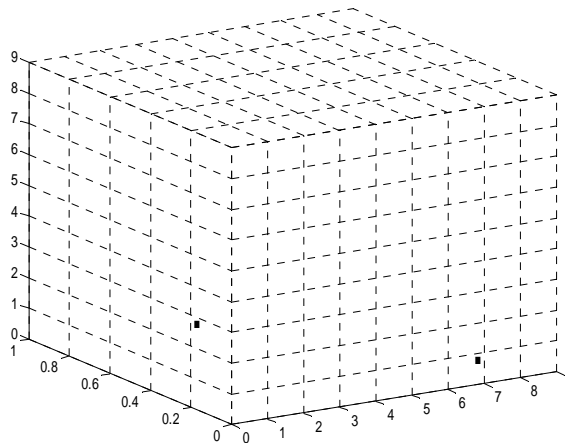


Figure (2.19) Two 3-dimensional vectors in space

Following on from the above formula, we can expand it for increasing dimensionality 4, 5, n . For a four-dimensional space in the same range $0 \dots 9$ the maximum number of possible vectors is 10^4 or $10 \times 10 \times 10 \times 10 = 10000$, for 10 dimensional space is $10^{10} = 10,000,000,000$ vectors and so on.

The conclusion is that a fixed number of data vectors occupies proportionately less and

less of the data space with growing dimensionality and, in terms of density of data in a space, this means that the data space size becomes sparsely populated by data vectors and the shape of the manifold is increasingly poorly defined. This leads to the following question: Why this the rapid increase of vector space with dimensionality occurs? In practice, the answer to that question would be that this rapid increase in data space size with dimensionality is often referred to as the “curse of dimensionality” (e.g. Moisl, 2009a, 2009b, 2011, 2012, 2015; Maguire & McMahon, 2011; Lee & Verleysen, 2007; Steinbach, 2004; Köppen, 2000; Bellman 1961), and it is a problem in many subject areas of science and engineering. For cluster analysis it is a problem because, the higher the dimensionality, the more difficult it becomes to fill the space or part of the space to characterize the manifold and thus to achieve a mathematically sound and reliable analysis. The explanation to this problem can be justified with reference to the ratio of actual to possible vectors in the space. In general, for a data set of fixed size D , the ratio of actual to possible vectors in the space is D/r^n , where D is the dimensionality and r^n is the number of vectors that can take integer values in a given range. To see how and why, suppose that we want to analyse, say, 10 texts in terms of their usage frequency of a single function word ‘as’; assume also that this function word is rarely used, so a range of 1...10 is sufficient. It is highly likely that the ratio of actual to possible vectors in the space is $10/10=1$, that is, the vector occupies the whole of the available space. If one analyses the 10 texts in terms of their usage frequency of 2 function words ‘few’ and ‘of’, also in the range of 1...10. It is quite likely that the ratio of actual to possible vectors in the space is $10 / (10 \times 10) = 0.1$, that is, that some spaces will be empty since the vectors occupy 10% of the available data space. In the same way, if one analyses the 10 texts in terms of 3 function word usage frequencies, the ratio of actual to possible vectors is $10/1000=0.01$ or 1% of the data space. In the 8-dimensional case it is $10/100000000$, or 0.0000001% and so on for increasing dimensionality. Dimensionality has a large effect on the ratio of actual to possible data points in the space: while the dimensionality rises, the ratio of actual to possible vectors in vector space falls at an exponential rate.

It is obvious from just looking at the successive percentages above what the overall indication is. As there are far more empty cells or locations in the space than vectors, the data space becomes very sparsely populated by vectors (i.e. the space usually involves empty slots where vectors would go). Metaphorically speaking, the constituent data points representing the plotted vectors in the manifold will be lost in the n - dimensional space, as in Figure (2.20) below.

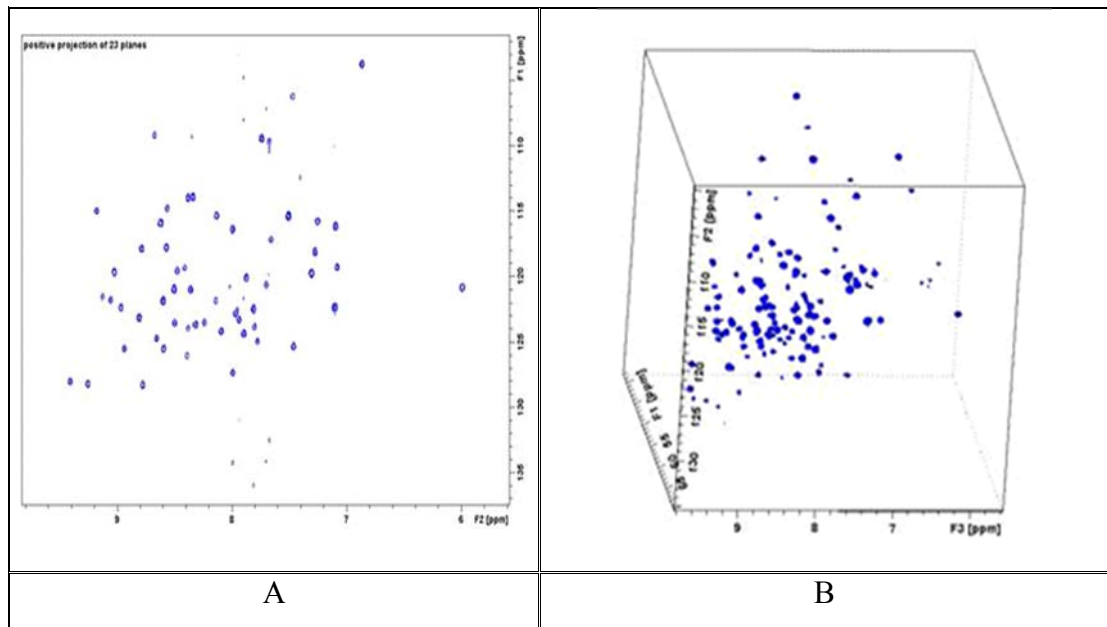


Figure (2.20) Sparse data in the space: 10% of data in 2-dimensional space (A) and in 3-dimensional space (B)

Getting enough or a fixed number of data vectors, therefore, becomes a serious problem even at relatively low dimensionalities and is usually difficult or even intractable because the dimensionality grows. What about gathering more data vectors as an alternative to fill the cells and improve the occupancy of vectors in the space? Assume that a given manifold needs 50% occupancy of the data space to be adequate for the manifold definition and to achieve that occupancy for the 2-dimensional space in a range of 0...9 one would need 50 vectors, 500 vectors for 3-dimensional space, 5000 for 4-dimensional space, and 5,000,000,000 for the 10-dimensional one. This may or may not be possible. What would the number of vectors be for dimensionalities higher than 10?. After all, the alternative of adding data vectors to improve a sparse manifold in most cases is not always practically possible.

Another question needs to be addressed here is that why all this talk about a manifold and dimensionality? In a practical setting, the answer to that question would be that clustering, dimensionality reduction, and a manifold have interesting relationships that have a particular relevance to the present application:

1. Any lexical frequency data matrix derived from natural language corpus will, in general, be very sparse on account of the large number of very infrequent lexical type

stylistic features (Stamatatos, 2009; Forsyth & Holmes, 1996; Holmes, 1994; Herdan, 1964); in the case of Coleridge's data matrix, there are 53 vectors in a 265-dimensional space, which is very sparse indeed. These infrequent features do not contribute to revealing the clustering patterns or they may even obscure the hidden clusters because of curse of dimensionality; details of which are given in the course of discussion.

2. Most popular cluster analysis methods (e.g. hierarchical cluster analysis, principal components analysis, multidimensional scaling, self-organizing map, Isomap) group vectors on the basis of their relative distances from one another in a vector space. Given the aim here is to use cluster analysis, the problem is that the distances between pairs of vectors in the space approach regularity due to the growth in dimensionality and therefore it becomes less and less possible to cluster the texts reliably.

This is what dimensionality does to cluster analysis, and it does so in the following ways: When dimensionality grows the distance between any two vectors in multidimensional vector space become increasingly close or similar to each other and this increase in closeness or similarity occurs very rapidly at relatively low dimensionality and then stop increasing or reduced. This means that it quickly becomes increasingly difficult to discriminate vectors from one another on the basis of distance among them, as in Figure (2.21). However, this phenomenon, where the vectors are no longer dissimilar or the distance between any two vectors in the dimensional space are the same for all vectors or close, is called 'concentration of distances' where the discrimination of 'nearest and farthest point/neighbour' in particular becomes meaningless (Moisl, 2015; Kab'an, 2012; Durrant and Kab'an, 2009; Clarke, et al. 2008; Beyer, et al. 1999; Saw, et al., 1984).

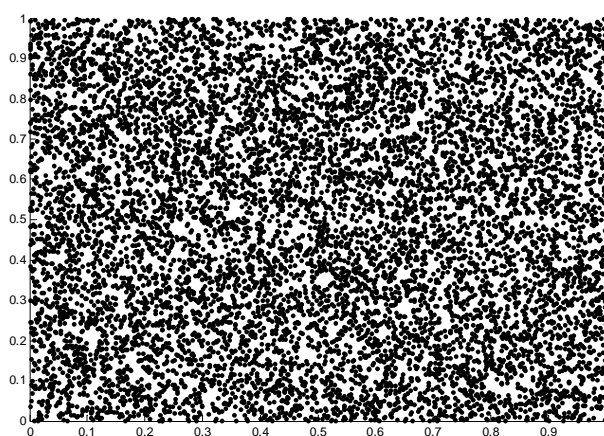


Figure (2.21) Concentration of distances among vectors in space

In the associated information retrieval and data mining literature (e.g. Kantardzic, 2011;

Cios, et al., 2007; Wang, et al., 2005), proximity (i.e. similarity or dissimilarity) between vectors in a space is articulated simply as the ‘nearest or farthest point/neighbour’ search. For the nearest neighbour, if cluster A is the set of vectors v_1, v_2, \dots, v_m and cluster B is v_1, v_2, \dots, v_m , the nearest distance between clusters A and B is $D(A, B) =$ (the shortest distance) minimum d_{ij} : where vector v_i is in cluster A and vector v_j is cluster B and d_{ij} is the Euclidean distance between v_i and v_j . For the farthest neighbour, if cluster A is the set of vectors v_1, v_2, \dots, v_m and cluster B is v_1, v_2, \dots, v_m , the farthest distance between clusters A and B is $D(A, B) =$ (the longest distance) maximum d_{ij} : where vector v_i is in cluster A and vector v_j is cluster B and d_{ij} is the Euclidean distance between v_i and v_j (Moisl, 2015; Chen, 2012). But when dimensionality grows, however, this straightforward approach becomes increasingly unreliable because “under certain broad conditions . . . as dimensionality increases, the distance to the nearest neighbour approaches the distance to the farthest neighbour. In other words, the contrast in differences to different data points becomes non-existent” (Beyer et al. 1999, cited in Moisl, 2015:75); on this see further: Francois, Wertz, and Verleysen (2007); Steinbach, Ertöz, and Kumar (2004); Aggarwal, Hinneburg, and Keim (2001); Korn, Pagel, and Faloutsos (2001); and François, Hinneburg, Aggarwal, and Keim (2000). This effect can, moreover, appear for dimensionalities as low as 10–15 (Beyer et al. 1999).

- Dimensionality reduction methods:

One solution to these problems of high-dimensional data in vector space is either to use the data as it is, which means that the analysis will be run to a poor degree of accuracy from which unreliable outcomes will be obtained, or to reduce the sparsity (Moisl, 2009a, 2015; Verleysen, 2003, 2008; Lee & Verleysen 2007; Priddy & Keller, 2005; Verleysen et al., 2003; Pyle, 1999; Bishop, 1995; Scott & Thompson, 1983). The present discussion adopts the second course; the remainder of this section addresses some ways of achieving dimensionality reduction.

Dimensionality reduction has been studied extensively (e.g. Moisl, 2015; Crain et al., 2012; Koppel et al., 2009; Juola, 2008; Verleysen 2008; Lee & Verleysen 2007; Priddy & Keller, 2005; Verleysen, 2003; Verleysen et al. 2003; Belew, 2000; Bishop, 1995; van Rijsbergen 1979; Salton & McGill 1983; Luhn, 1957, 1958), discussed in Salton and McGill (1983: 60-63), and in that literature a large number of data reduction methods have been proposed. The following discussion selects a few which seem particularly

appropriate to the present application.

Dimensionality reduction methods can be divided into two major types: variable selection and variable extraction. The first of these tries to identify a subset of the more important researcher-defined variables and to remove the remainder from the analysis (given some definition of importance) without losing too much information, thereby achieving dimensionality reduction. The second replaces the set of researcher-defined variables with a smaller set of variables which reduces dimensionality but captures most of the variability in the original set. The second of these often achieves a greater degree of dimensionality reduction, but at a cost: the newly-defined variables are generated by mathematical procedures, and their meaning relative to the research domain is typically difficult to determine reliably. This study will need to retain the meaningfulness of the variables it uses, as will be seen, and as such variable extraction is not used. For further information on variable extraction see, for example, Moisl (2015), Martinez, Martinez, and Solka (2011), Lee & Verleysen (2007), Camastra (2003), Verleysen (2003), and Jain and Dubes (1988). The remainder of this section describes some variable selection methods.

Given that variable selection methods aim to select a subset of the more important variables, a well-defined criterion of importance is fundamental. Two of the most often used ones in the literature are frequency and variability, and these are described below. Others, such as term frequency-inverse document frequency (TF-IDF) and measures of nonrandomness, are also available, but these gave results similar to those based on frequency and variability in the analyses described later in the discussion, and the additional complexity associated with them was therefore felt not to justify their inclusion; for further information on these see (Moisl, 2015: 78-114).

a. Frequency:

Frequency is the simplest criterion for selecting features from a data matrix: those variables which occur most often in the research domain — in the present domain, words in text— are judged to be the most important, and those which occur least often are taken to be least important and can therefore be discarded (Thomas et al., 2004; Culpeper, 2002; Holmes, 1992, 1994, 1998; McEnery & Wilson, 1996; Ide & Walker, 1993; Ide, 1989; Burrows, 1987; Beardsworth, 1980; Iker, 1974; Enkvist, 1964, 1973; de Sola Pool, 1959; Saporta & Sebeok, 1959). 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. Here is as an example, based on Williamson (2009): Suppose we count the number of words in a text and find that there are 87 tokens of 62 types. Suppose also we find that more than half of the types (51) are hapax legomena (i.e. words occurring once), four types are hapax dislegomena (words occurring twice), and four words occur three times, the word ‘we’ occurs 6 times, and the word ‘them’ occurs 5 times. In such a case, the conclusion would be that ‘we’ and ‘them’ are frequent words and therefore must be taken into consideration when attempting to analyse that text, whereas the other types (e.g. hapax legomena) are infrequent words (since they tell little or nothing about that text) and can be taken as just random noise that adversely affects the results (Kaufman & Rousseeuw, 1990). The moral of the example is that word frequency is fundamental in authorship attribution studies and lexical statistics, and more detailed information about this can be found in, for example, Baayen (2001).

To select variables based on frequency, given an $m \times n$ frequency data matrix D ; the value at D_{ij} is the number of times variable j , for $j=1\dots n$, occurs in text i , for $i=1\dots m$. The frequency of occurrence of variable j across the entire corpus of texts is then:

$$freq(F_j) = \sum_{i=1..m} F_{i,j}$$

Frequencies of for all the columns data matrix D are calculated, the variables are sorted in descending order of frequency, the most frequent variables are selected, and the less frequent variables are eliminated from D . Substantial dimensionality reduction can be achieved by applying this criterion to a data matrix D .

b. Variability:

Variability refers to the amount of variation in the values that a variable takes. Any variable x is an interpretation of some aspect of the physical world, and a value assigned to x is a measurement of the world in terms of that interpretation. If x is to describe the ages of people, it can take different values for different persons or for the same person at different times. Unless all people are exactly the same age, or the age of the same person is fixed, the values which x takes will vary substantially, and can, therefore, contribute to

the distinction of people from one another, or of the age of same person at different times (i.e. the more different people groups one tests, the more variation one will see in the ages). This possibility of variability in the values assigned to variable x gives it its descriptive utility: an identical value for x tells that what x stands for in the real world does not change, moderate variability in the value tells that aspect of the world changes only a little, and widely differing values tells that it changes substantially. In general, therefore, the possibility of variability in the values assigned to variables is necessary to the ability of variables to describe objects and thereby to represent reality.

Clustering of texts or of anything else depends on there being variability in their characteristics; identical texts having the same stylistic descriptors cannot be meaningfully clustered. When the texts to be clustered are described by variables, then the variables are only useful for the purpose if there is significant variation in the values that they take. If, for example, a large number of people were described by their weights or heights, we would expect there to be logically substantial variation in values for each of them, and any cluster analysis method could legitimately be used to cluster them. On the other hand, if a large number of people were described by variables like ‘eyes’, ‘noses’, and ‘legs’, there would be almost no or little variation or high correlation with other features, since, with very few exceptions, everyone has two eyes and a nose, and clustering based on these variables would be effectively useless. In any clustering application, therefore, one is looking for variables with substantial variation in their values, and can ignore variables with little or no variation. Variables with no or little variation should be removed from data matrix as they contain little information and complicate cluster analysis by making the data higher-dimensionality than it needs to be (Moisl, 2009a, 2009b).

Mathematically, the degree of variation in the values of a variable is described by its variance (Moisl, 2015; Pyle 1999). We begin with the mean or average of variable values. Say a variable x represented as a vector of 10 numerical values across some range. The mean, or average, is a measure of the central tendency (or, more commonly, a measure of a typical value) of a variable x :

X	5	10	15	20	25	30	35	40	45	50
-----	---	----	----	----	----	----	----	----	----	----

Table (2.4) An example of a mean

Visual inspection suggests that the value at the centre of vector elements is around 25 or 30. A more precise inspection can be given by adding all vector values and then dividing by their numbers: $5 + 10 + \dots + 50 = 275/10 = 27.5$.

Mathematically, this can be expressed as:

$$mean(x) = (\sum_{i=1..n} (x_i)) / n$$

The mean often hides important information about the distribution of values for a given variable. That is, the mean works well when most of the individual values are close to the mean. But if the values vary greatly, the mean may take a typical value and could be misleading (Ehrenberg, 1982).

As an example, the following table shows the frequency counts of two variables (X and Y) occurring in the corresponding ten texts (1.....10):

X	40	58	92	31	27	85	67	77	73	32	Mean	60
Y	55	62	56	46	59	58	57	54	58	59	Mean	60

Table (2.5) An example of 10 values for two variables *X* and *Y*

The means for variables *X* and *Y* are identical, but the variations of frequency counts across the texts differ significantly: variable *X* is able to demonstrate both high and low frequency values, and variable *Y* is relatively constant. Knowing only the mean one could not make the distinction between *X* and *Y*; both the mean and some indication of the spread of frequency values across are required. In the above example, where the number of variables is few, visual inspection is sufficient (Moisl, 2009a, 2009b), but what if there are a large number of values with a long range of word frequencies? Visual inspection quickly fails; such assessment must be less dependent on visual inspection and some quantitative measure that summarizes the spread of frequency values is required. That measure is variance.

The variance of a set of variable values is the average deviation of those values from their mean (Moisl, 2011; Rencher, 2002). Assume a set of *n* values $\{x_1, x_2 \dots x_n\}$ assigned to a

variable x . The mean of these values μ is $(x_1 + x_2 + \dots + x_n) / n$. The amount by which any given value x_i differs from μ is then $x_i - \mu$. The mean difference from μ across all values is therefore $\sum_{i=1..n} (x_i - \mu) / n$. This mean difference of variable values from their mean almost but not quite corresponds to the definition of variance. One more step is necessary, and it is technical rather than conceptual. Because μ is an average, some of the variable values will be greater than μ , and some will be less. Consequently, some of the differences $(x_i - \mu)$ will be positive and some negative. When all the $(x_i - \mu)$ are added up, as above, they will cancel each other out. To prevent this, the $(x_i - \mu)$ are squared. The standard definition of variance for n values $\{x_1, x_2, \dots, x_n\}$ assigned to a variable x , therefore, is:

$$v = \left(\sum_{i=1..n} (x_i - \mu)^2 \right) / n$$

Thus, in Table (2.5), the variance of X is $((40-60)^2 + (58 - 60)^2 + (92 - 60)^2 \dots + 30-60)^2) / 10 = 316.31$. Doing the same calculation for variable Y , the variance works out as 10.00. Comparing the two values, it is clear that the variability in X 's frequency values is much greater than Y 's, the larger the value of the variance, the more the numbers differ from the mean and the smaller the value, the less they differ.

Interpretation of variance is not as straightforward as it appears to be. In the above example, what do the magnitudes mean in absolute terms? When several variances are compared, the relativities of the magnitudes reflect degrees of variation, but what if one is trying to interpret a single variance without reference to others? What, in absolute terms, does 316.31 indicate about the amount of variation in X ? Is it a large variation or a small one? The problem is that the squares quantities are not readily interpreted in terms of the original units of measurement. To recover the original units, it is only necessary to take the square root of the variance. Doing this for the above variances, the square root of 316.31 is 17.78, and for 10.00 it is 3.16. The interpretation is that, for variable x , the average divergence of frequency counts to either side of the mean is 17.78 and for Y , it is 3.16, the reasonableness of this a quick glance at the range of frequency counts will confirm. The square root of variance is the standard deviation, and it gives a measure of the average deviation from the mean of variable values in terms of the original variable range. Variation expressed in terms of the original variable range is more intuitively meaningful than in terms of variances, which are just numbers whose only interpretable significance is the difference in magnitude. Because of their interpretability relative to the

variable values on which they are based, standard deviations are most often used in preference to variances in quantifying the spread of values across a range.

Given a data matrix M in which the row vectors are data items of interest and the column vectors are lexical type variables describing the texts of query, and also that the aim is to cluster analyze these texts on the basis of the differences among them, the application of variance/standard deviation to dimensionality reduction is straightforward: calculate and plot the variances of the columns and, if any have variability which is low in relation to that of the others, remove them on the grounds that they contribute little to differentiation of the texts, and decide on a threshold selection (the set of retained variables from each column of the data matrix) (Moisl, 2015, 2009a, 2009b; Milton & Arnold, 2003; Pyle, 1999).

There is, however, a caution in using variance / standard deviation as a selection criterion. When the variables are measured on different scales, variance in itself presents a problem as the measurements of variables' relative variations based on their variances can be misleading. If one variable has a much wider range than others then this variable will tend to dominate. For example, if distance measurements had been taken between a number of different things, the range in centimetres or meters of lengths would be much wider than the range in kilometres or miles, a difference of 10 kilometres could being a difference 1000 meters, say. The distinction between absolute and intrinsic variability (also known as the between-cluster variability) has particular relevance for understanding the problem of disparity in variable scale. Absolute variability is the amount of variation in values expressed in terms of the scale on which those values are measured, and is measured by standard deviation. On the other hand, intrinsic variability refers to the amount of variation expressed independently of scale, and is measured by coefficient of variation (Moisl, 2010, 2015; Everitt, 2011; Chu et al., 2009; Boslaugh & Watters, 2008; Gnanandesikan et al., 1995; Milligan & Cooper, 1988; Anderberg, 1973). In cases where variables are measured on different scales and the range of value differs widely from one variable to another, the comparison of the standard deviations of a set of variables under consideration therefore carry different amount of information or a scale dependent assessment of their variations. As the magnitude of a variable's values has strong effect on the variable's standard deviation, a variable with a relatively lower intrinsic variability but relatively larger values can dominate or influence the results than one with relatively higher intrinsic variability but relatively smaller values. In cases where such disparity of

variable scale exists, coefficient of variation measure of intrinsic variability is normally used as a criterion for variable selection (Moisl, 2010, 2015; Chu et al., 2009; Gnanandesikan et al., 1995; Milligan & Cooper, 1988).

The advantage of using variability as a selection criterion is, of course, that it is mathematically easy to understand and straightforward to apply: high variance variables are important in distinguishing between texts in a collection, and low variance ones are not.

2.2.3.4 Data Analysis:

It was noted earlier in this chapter that testing of the hypothesis under discussion would be based on finding structure in high-dimensional data space and then using that structure to attempt to falsify the hypothesis. One way of finding structure in high-dimensional data is cluster analysis, and that is the approach taken here. Cluster analysis includes an extensive variety of mathematically-based methods; for an overview see (Moisl, 2015). The present section first introduces the concept of clustering and then describes the selection of clustering methods used in subsequent chapters.

a. What is a cluster?

The human perceptual system is optimized to detect patterning in the environment (Moisl, 2015: 153). Figures (2.22a-2.22d), taken from (Moisl, 2015: 154), are plots of two-dimensional data representing objects in some real-world domain of interest. Any observer can identify the presence or absence of patterning in them, and from that can infer the presence or absence of structure in the domain they represent.

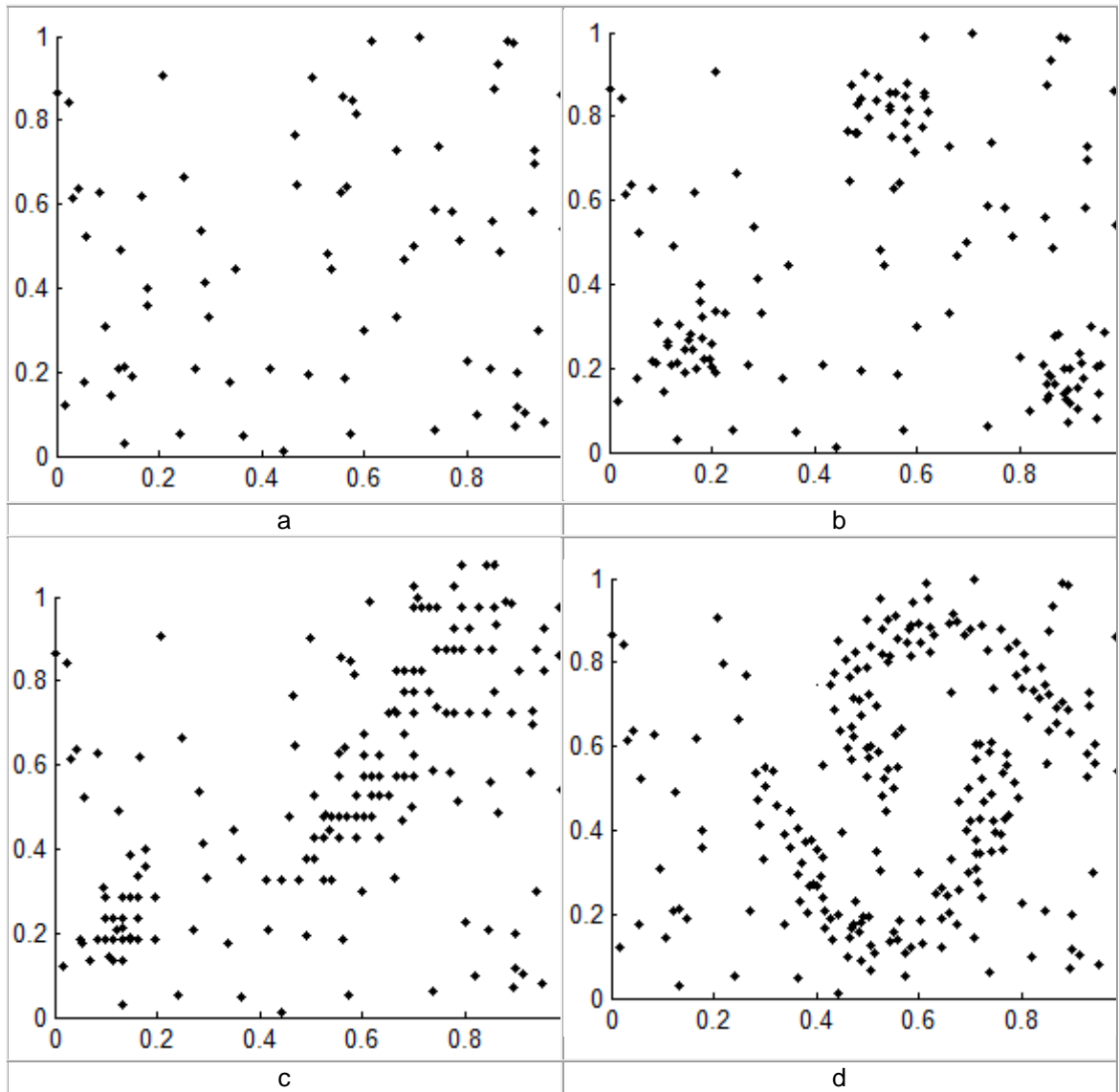


Figure (2.22) Scatter plots of 2-dimensional data

Figure (2.22a) is just a random scatter of points, and indicates that the domain is unstructured: the objects in the domain have no discernible pattern of relationship to one another. Figure (2.22b) has three concentrations of points in a background of random scatter, and indicates that most of the objects in the domain fall into three groups. Figure (2.22c) has two concentrations of points, one large and one small, again against a random scatter, and indicates that most of the objects represented fall into two groups of unequal size, and analogously for (2.22d). These point-concentrations are clusters.

Direct perception provides an intuition for the concept of clustering, but is of limited use for data analysis, for two reasons (Moisl, 2015: 154). One reason is that interpretation of clusters is subjective, and more specifically is dependent on the researcher's knowledge of

the data domain, which can lead to biased conclusions. The other, more serious reason is that direct perception of clusters is limited to three dimensions, and is therefore not extendable to data of higher dimensionality; one might argue that this can be solved by reducing data dimensionality to three or fewer, but this assumes that such reduction is possible without losing essential domain information, and that assumption is not necessarily justified.

There is no obvious solution to the problem of subjectivity, but cluster analysis is not alone in this. All results from all methodologies in science are ultimately interpreted by humans, and there is no absolutely objective human interpretation; to do science is to interpret subjectively (e.g. Soffer, 1987; Nesterenko, 1979). The dimensionality problem, on the other hand, has a solution, at least in principle: formulate a mathematical definition of what a cluster is, and then design mathematical methods for identifying clusters in data with dimensionality greater than three relative to that definition. In practice, a generally agreed definition of what a cluster is has not yet been formulated. Moisl (2015: 155) notes that there are two main ways to conceptualize a cluster. The first is to conceptualize clustering as distance among objects in data space. The other is to conceptualize clustering as variation in the density of objects in the space. Moisl further quotes the following formulations of these two views from a standard textbook on cluster analysis (Jain & Dubes 1998: 1):

- “A cluster is an aggregation of points in the test space such that the distance between any two points in the cluster is less than the distance between any point in the cluster and any point not in it.”
- “Clusters may be described as connected regions of multi-dimensional space containing a relatively high density of points, separated from other such regions by a region containing a relatively low density of points”.

The account of data creation and transformation earlier in this discussion conceptualized data in terms of relative distance among points in vector space, and the distance-based formulation of clustering is consequently adopted in what follows. This is a purely practical choice: inclusion of density-based clustering methods would have extended the discussion substantially beyond what is expected of a PhD thesis. The choice of the distance over the density view of clustering is not intended to imply the superiority of the former; as the literature shows (Moisl, 2015: 155-6), the latter shows considerable

promise, and research subsequent to this dissertation may well apply it to the problem under discussion.

b. Clustering methods:

This section first describes the clustering methods used in the subsequent chapter — principal components analysis, multidimensional scaling, Isomap, the Self-Organizing Map, and hierarchical clustering— and then justifies the choice of these specific methods.

i. Principal Components Analysis:

Principal Components Analysis (PCA) is actually a dimensionality reduction method which the preceding discussion of that topic decided against using because it entailed redefinition of the variables used to describe the objects in the data domain, but it can also be used for clustering if the dimensionality is sufficiently reduced. The conceptual basis of PCA is elimination of variable redundancy. Selection of the set of variables to describe the objects in a research domain is at the discretion of the researcher, as noted earlier. It was also noted earlier that that selection in any given application is not necessarily optimal. Such non-optimality is manifested as redundancy among variables, that is, as overlap in the information which the variables provide; the variables 'Age' and 'Income' in the description of people, for example, are redundant because there is a correlation between them: in general, the older people are the more they earn, up to retirement at least. PCA aims to identify such redundancy in the researcher-defined set of variables and to replace them with a new and smaller set of non-redundant variables. Specifically, given a matrix of m data items described by n variables, principal components analysis is a technique for redescribing the m items in terms of k variables, where $k < n$, such that most of the variability in the original n variables is retained. When $k = 2$ or $k = 3$ the m data items can be plotted in two or three dimensional space and any clusters can thereby be directly perceived.

Using PCA as a clustering method for the rows of a given high-dimensional data matrix M implicitly assumes, of course, that there is redundancy in M . If not, reduction to two or three dimensions would lose essential information captured by the original set of variables, and the resulting clusters would be based on partial information, possibly leading to misleading results.

Redundancy among variables is determined by measuring the similarity among the column vectors in the data matrix. There are various such measures:

- Distance: The values in an n -dimensional vector or the coordinates of its location in n -dimensional space. The similarity of any two vectors in the space is consequently reflected in the distance between them: vectors with very similar values are close together, and vectors with very different values far apart. By calculating the distances between all unique pairings of column vectors in a data matrix, the degrees of similarity and therefore of redundancy between them can be determined.
- Angle: The angle between a pair of vectors in a vector space reflects the distance between them, assuming that the vectors are of equal length. The degrees of similarity and therefore of redundancy between all unique pairings of column vectors in a data matrix can be found by calculating the cosines of the angles between them: the smaller the cosine the larger the distance between column vectors, and therefore the smaller the redundancy.
- Covariance / Correlation: In probability theory two events A and B are independent if the occurrence of A has no effect on the probability of B occurring, or vice versa, and dependent otherwise. Given two variables x and y and an ordered sequence of n observations at times $t_1, t_2, t_3 \dots t_n$ for each, if the measured value for x at time t_i (for $i = 1..n$) has no predictive effect on what the value of y will be at time t_i , then the variables are independent, or, failing that condition, dependent. In statistics, variables that are dependent are said to be associated, and the degree of association is the degree to which they depart from independence. Statistics provides various measures of association, the most often used of which is Pearson's Correlation Coefficient, defined as:

$$P_{corr} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

where $P_{corr}(x,y)$ is the Pearson Correlation Coefficient of x and y , σ_x and σ_y are the standard deviations of x and y respectively, and $\text{cov}(x,y)$ is the covariance of x and y , defined as:

$$\rho(x, y) = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

where μ_x and μ_y are the means of x and y respectively.

In principle, either covariance or Pearson Correlation can be used to measure association. Pearson Correlation has the advantage of being more easily interpretable than covariance because it is always in the range $-1...1$, whereas covariance is dependent on the scales on which the data variables are measured. If all the variables are measured on the same scale, however, this doesn't matter, and the choice between covariance and correlation is neutral.

As for distance and angle, the covariances or the Pearson Correlation Coefficients for all unique pairings of column vectors in a data matrix can be calculated and the degree of redundancy of each determined: the greater the correlation coefficient, the greater the redundancy.

Given an n -dimensional data matrix containing some degree of redundancy, PCA replaces the n variables with a smaller set of k uncorrelated variables called principal components which retain most of the variance in the original variables, thereby reducing the dimensionality of the data with only a relatively small loss of information. It does this by projecting the n -dimensional data into the k -dimensional vector space, using a two-step process: the first step identifies the reduced-dimensionality space, and the second projects the original data into it.

We now begin by looking at the standard two-dimensional Cartesian basis, where there are two dimensional vectors, one with dimension x and one with y . A plot showing the relationship between these two vectors might look like this:

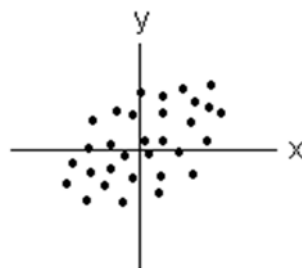


Figure (2.23) Two-dimensional data distribution with orthogonal basis

It is mathematically possible to rotate the basis to find a new X-Y orthogonal basis for this distribution of points in such a way that each axis is a best fit for the main directions of variability among the points of vectors. The most important thing to bear in mind is that the rotation of X-Y axis, one or the other of them, need to be orthogonal to one another (90 degrees or uncorrelated with one another). However, the line of best fit X' is

drawn through the points, and the line of second-best fit Y' along in such a way that Y' is orthogonal to X' . Thus in some sense, the line of best fit X' goes through the maximum variability of each point to that line since it is as close to all points as possible and so is the line of second-best fit Y' , as shown in figure (2.24):

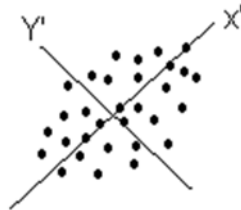


Figure (2.24): Alternative orthogonal basis for data

The vector axes are then reframed or rotated in other direction relative to the new X-Y orthogonal basis. In the frame of dimensionality reduction this doesn't get us any further, since it simply reframes the original data in two dimensional spaces in respect of a different X-Y orthogonal basis, i.e. nothing has been done to data itself, we are just looking at it from a different angle. As an example, consider the following distribution of a group of points in which the vectors are highly correlated:



Figure (2.25) Highly correlated two-dimensional vectors distribution with orthogonal basis

If the orthogonal lines of best and second-best fit are sketched here, the points projected on them will look like this

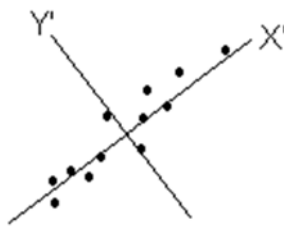


Figure (2.26) Alternative orthogonal basis for vectors

then it is clear that on Y' the points are not very spread out here, therefore they do not have a large variance. On X' the points are very spread out, they have a large variance. In other words, X' best describes almost all the variability among the points of vectors, and Y' describes only a small amount. Now, if Y' is simply ignored, then the points of vectors can be reframed in 1 rather than the original 2 dimensional spaces with minimum loss of information, and the data dimensionality has been reduced.

This idea extends to any dimensionality, i.e. higher dimensional spaces. However, with more than three dimensions, though the visualisation of data points usually becomes difficult or impossible. In the three-dimensional case (2.27a), the first two dimensions Z' and Y' are sufficient to represent the vectors, achieving a dimensionality reduction of 3 to 2, and in case (2.27b) the dimensionality can be reduced to 1 by using only the Z' dimension.

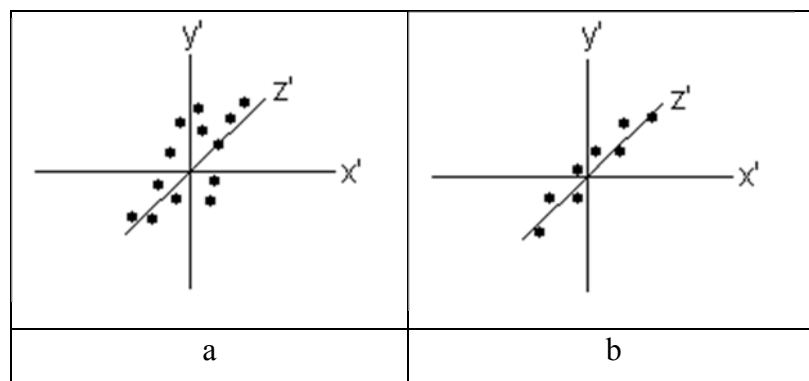


Figure (2.27): Three-dimensional data distribution with orthogonal basis

Relative to points of vectors in n -dimensional space, then, there are three main steps to find the principal components of a data matrix:

- Find an orthogonal basis for any given n -dimensional data matrix D without changing

either the variable variances or their covariance.

- Construct a few basis vectors (i.e. we call them principal components) for data matrix D , in such a way that each axis is the least-squares best fit to one of the n directions of maximum of variation in D .
- Remove the axes along which that have relatively little variation, leaving an m -dimensional basis for D , where $m < n$. Once again, the data matrix is D , an $m \times n$ matrix, where m is the number of variables and n is the number of texts.
- Project the original n -dimensional data D onto the reduced m -dimensional space, which yields a data set D' that is dimensionality-reduced but still has the property of maximum variation in D , that is, the total combined variance of all vectors.

What is required now is a mathematical procedure to perform these steps, and PCA provides it. The discussion of PCA proceeds in the following stages:

1. Construction of a similarity matrix:

PCA is based not on the given data matrix D , but on a matrix C of similarities between the column vectors of D , using one of the measures of similarity described above; covariance is used here. Given a set of variables in n -dimensional space, PCA calculates their covariances and saves them in C , where C is an $n \times n$ square matrix in which both the rows i and the columns j (for $i, j = 1 \dots n$) represent the variables in the original data, and cell C_{ij} represents the covariance between variable i and variable j , that is, the covariance of column i and variable column j of D . Consider, for example, the following covariance matrix from (Moisl, 2015: 108) abstracted from a data matrix which contains the frequencies of six phonetic segments for a set of speakers.

	[ə ₁]	[ə ₂]	[ɔ:]	[ə ₃]	[l]	[eɪ]
[ə ₁]	19.20	5.67	-36.34	-10.96	-9.03	-27.16
[ə ₂]	5.67	4.57	3.05	-1.93	6.89	-20.40
[ɔ:]	-36.34	3.05	643.77	-36.99	-78.14	-18.39
[ə ₃]	-10.96	-1.93	-36.99	242.91	96.09	-62.66
[l]	-9.03	6.89	-78.14	96.09	205.14	-115.87
[eɪ]	-27.16	-20.40	-18.39	-62.66	-115.87	190.64

Figure (2.28) $n \times n$ covariance matrix of 6 phonetic segments for DMC

This matrix says that the covariance of phonetic segment [ə1] and phonetic segment [ɔ:] is -36.34, of phonetic segment [ə3] and phonetic segment [eI] is -62.66, between phonetic segment [eI] and phonetic segment [ə2] -20.40, and so on.

2. Construction of an orthogonal basis for the covariance matrix

An n -dimensional orthogonal basis for the $n \times n$ covariance matrix C is constructed such that every vector $V (v_1, v_2, v_3, \dots, v_n)$ in n -dimensional space is a linear combination of the standard orthogonal basis having the least-squares best fit to one of the n directions in C :

- The first basis vector v_1 is the vector of the best least-squares along the direction of maximum variation (among all linear combination) in C .
- The second basis vector v_2 is the next best least-squares along direction of maximum variation in C ; it is orthogonal to v_1 (i.e. uncorrelated or the correlation is 0) that accounts for as much of the remaining variation as possible.
- The third basis vector v_3 is another line of best fit along the third direction of maximum variance in C and is orthogonal to both v_1 and v_2 .
- All subsequent basis vectors v_4, \dots, v_n have this same property: each axis is orthogonal to or not correlated with other or the previous principal axis such that each is orthogonal to all other v_i for $i=1..n$ and has as much of the maximum or the remaining variation as possible.
- Each vector $V (v_1, v_2, v_3, \dots, v_n)$ is a principal component of C ; in each successive stage of constructing these components, we calculate the variance along each component, store the total components as the column of $n \times n$ matrix in descending order of the magnitude (i.e. covariance) they represent, and search for the next direction of maximum variation in D .

The standard procedure of constructing such orthogonal basis is to calculate the n eigenvectors and eigenvalues of the n eigenvectors of the covariance matrix C

$$[E_1 \ E_2] = \text{eig}(C)$$

where E_1 is a square matrix of the same dimensionality as C whose columns are the eigenvectors of C , E_2 is a square matrix of the same dimensionality as C whose positive diagonal contains the eigenvalues corresponding to the eigenvectors in E_1 , and eig is a

function that calculates E_1 and E_2 from C . Calculation of eigenvectors is a fairly complex matter, and a description of it is not needed here because the details are not germane to the discussion. Most linear algebra textbooks provide accessible accounts; see for example (Lay, 2010). The main thing is to realize that the eigenvectors of the covariance matrix constitute an orthogonal basis for it.

3. Selection of dimensions:

The orthogonal basis for an n -dimensional set of vectors is n -dimensional; applied to the $n \times n$ covariance matrix C , there are n eigenvectors. To perform dimensionality reduction, a procedure has to be found of removing the axes that define the direction of relatively little variation. The eigenvalue matrix gives the criterion for this: eigenvectors and eigenvalues appear in pairs in which each eigenvector has a corresponding eigenvalue. Say we have two variables in a 2 dimensional space, therefore there are 2 eigenvectors and values, for 3 variables in a 3-dimensional space, there are 3 eigenvectors and values, and so on to any dimensionality. An eigenvector is a direction of the line (e.g. vertical, horizontal, 45 degrees, etc) while an eigenvalue is a number indicates how much variation there is between and among variables in that direction (i.e. how spread out the variables is on a given line). The eigenvalues are therefore sorted in descending order of the variance they represent, that is, they are ranked from the highest to the lowest, and all the eigenvectors whose eigenvalues are below some specified threshold can be removed, giving an $n \times m$ eigenvector matrix E for C , where $m < n$. Selection of an appropriate threshold is discussed below (Moisl, 2015; Gaborski, 2014; Dallas, 2013; Singh, 2012; Richardson, 2009; and Annas, et al., 2007).

4. Projection into m -dimensional space:

Once the reduced-dimensionality eigenvector matrix E matrix has been found, it is used to project the original n -dimensional data set D into the reduced m -dimensional space, giving a new $n \times m$ dimensional data matrix D_{reduced} that still has most of the variation in D . This is calculated by the multiplication of the original n -dimensional matrix D^T by the reduced-dimensionality eigenvector matrix E^T_{reduced} , where T indicates matrix transposition, that is, create another matrix whereby the rows of the original matrix become columns and the column rows. This multiplication is defined by the following equation:

$$D^T_{reduced} = E^T_{reduced} \times D \text{ matrix}^T$$

When considering the application of PCA for dimensionality reduction, a number of computational questions arise (Moisl, 2015: 111-4):

- The suitability of the original data for analysis:

The original data needs to be mean-centred prior to generation of the covariance matrix. That is, for PCA to work properly the first step is to centre the data on zero; the mean must be subtracted from all the data dimensions where the mean subtracted is the average across each dimension. So, all v_1 values have the mean for v_1 subtracted from them, all the v_2 values the mean for v_2 , and so on. This produces a data set whose mean is zero (Moisl, 2015).

- Covariance or correlation matrix:

PCA can be calculated by using either a covariance function generated from a covariance matrix or a correlation function from a correlation matrix. If variables vary in scale, that is, do not have the same units of measurement, a correlation matrix is better. Otherwise, when variables have the same units of measurement as here measured on the same scale, the covariance matrix can be used.

- Selection of dimensionality:

When PCA is used for dimensionality reduction, the optimal number of components has to be selected, where optimality is the best balance between reduction and retention of variance from the original set of variables. Various methods for doing this exist (Moisl, 2015: 111-4), but for present purposes the required dimensionality is known in advance: to permit plotting in two or three dimensions, the maximum number of components is three.

- Variable (dimension) interpretation:

For any given data matrix, the variables generally have labels that are semantically important to the researcher in the sense that they describe aspects of the research field

considered to be relevant. Since PCA describes a new set of variables, these labels are not any more useful for the column vectors of the dimensionality-reduced matrix, and the values for them are self-evidently not interpretable as the frequencies of the original data since some of them are negative. Where, however, the aim is simply to reduce dimensionality for clustering, as here, the new variables do not require semantic interpretation, and as such this is not a problem.

For more on PCA, see, for example, Moisl (2015), Gaborski (2014), Dallas (2013), Jamak, et al. (2012), Singh (2012), Hair et al. (2010), Richardson (2009), Annas, et al. (2007), Jackson (2003), Jolliffe (2002), Rencher, (2002), Everitt & Dunn (2001), Tabachnik & Fidell (2001), Bishop (1995), Grimm & Yarnold (1995), Rietveld & van Hout (1993), Woods et al., (1986).

ii. Multidimensional scaling:

Like PCA, 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 differs from PCA in that, whereas PCA 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.

Given an $m \times m$ proximity matrix P derived from an $m \times n$ data matrix D using one of the distance measures described earlier, MDS finds an $m \times k$ reduced-dimensionality representation of D , where k is a user-specified parameter. MDS is not a single method but family variants. In the MDS literature (e.g. Moisl, 2015; Lee & Verleysen, 2007; Borg & Groenen, 2005; Wickelmaier, 2003) the distinction is usually made between the so-called classical MDS method and its variant metric least squares MDS, also known as nonmetric MDS. Classical MDS requires that the proximity measure on which it is to operate be Euclidean distance. Given an $m \times n$ data matrix D , therefore, the first step is to measure the $m \times m$ Euclidean distance matrix E for D . A simplified view of how the method works is as follows:

- We find mean-centred E by calculating the mean value for each row E_i (for $i = 1 \dots n$) and subtracting the mean from each value in E_i .
- We calculate an $m \times m$ matrix S each of whose values $S_{i,j}$ is the inner product of rows E_i and E_j , where the inner product is the sum of the product of the corresponding elements as described earlier in the discussion of vector space basis and the T superscript denotes transposition:

$$S_{i,j} = \sum_{k=1 \dots m} (E_{i,k} \times E_{j,k}^T)$$

- We calculate the eigenvectors and eigenvalues $E_1 E_2$ of S , as discussed above.
- We use the eigenvalues, as in PCA, to find the number of eigenvectors K ($k_1, k_2, k_3 \dots kn$) worth keeping.
- We project the original data matrix D into the reduced k -dimensional space, again as in PCA:

$$D_{reduced}^T = E_{reduced}^T \times D \text{ matrix}^T$$

This equation is very similar to PCA, it can in fact be shown that classical MDS and PCA are equivalent and give the same results (Moisl, 2015; Lee & Verleysen, 2007; Borg & Groenen, 1997, 2005), and are therefore just second or another solutions to a given problem. For this reason, a variant of classical MDS, known as Metric least squares or Nonmetric MDS, will be described here and used in the subsequent chapter. This alternative method extends the applicability of MDS beyond what PCA is able to perform, and provides the basis for additional dimensionality techniques more powerful than PCA and classical MDS. Metric least squares or nonmetric MDS aims to find a set of vectors in k dimensional space such that the matrix of distances among them corresponds as closely as possible to some function of the input matrix on the basis of a criterion called stress. More specifically, the problem of metric least squares MDS is how to find a mapping of row vectors (from higher-dimensional to lower-dimensional space) that minimizes the squared differences between the proximities or the distances between all distinct pairings of row vectors, that is, a configuration that minimizes the so called stress function to obtain the probable MDS map. Metric MDS works on distance measurement of proximity (similarity or dissimilarity) between pairs of row vectors. There exist various types of distance that this method can use. Euclidean distance is usually the first option for an

MDS space due to its simplicity of measurement and conceptual clarity, and is therefore used here. Given an $m \times m$ proximity matrix derived from an $m \times n$ data matrix D , metric least squares MDS creates an $m \times k$ representation matrix M' of an $m \times n$ matrix M by finding an M' for which the distances between all distinct pairs of data vectors i, j in M' are as close as possible to the proximities P_{ij} between equivalent data vectors of M , for $i, j = 1 \dots n$. The justification for this is that when the distance relationships in M and M' are adequately identical, M' is a sufficient reduced-dimensionality representation of M .

The projection f from M to M' could in principle be clearly expressed but is in practice estimated by the following iterative mathematical procedure:

1. We calculate the Euclidean distance matrix $D(M)$ for all distinct pairs (i, j) of the m row vectors of M , so that $\delta_{i, j} \in D(M)$ is the distance from row vector i to row vector j of M , for $i, j = 1 \dots n$.
2. We choose a dimensionality k and construct an $m \times k$ matrix M' in which m k -dimensional row vectors are randomly populated in the k -space.
3. We calculate the Euclidean distance matrix $D(M')$ for all distinct pairs i, j of the m row vectors of M' , so that $\delta_{i, j} \in D(M')$ is the distance from row vector i to row vector j of M' , for $i, j = 1 \dots n$.
4. In the last step, we compare the distance matrices $D(M)$ and $D(M')$ to decide on how close they are, where closeness is calculated on the basis of an objective function called a stress function. If the stress function arrives at a prearranged threshold of adequate closeness between $D(M)$ and $D(M')$, stop. Otherwise, it alters the values in the m row vectors of M' so that the distances between their new locations in the k -space come close to the equivalent ones in $D(M)$, and return to step (3) above.

“In simple terms”, searching for M' requires that we rotate its row vectors in the k -dimensional space until the distance relations between them become sufficiently close to those of the equivalent row vectors in M . The degree of equivalence between the distances among data vectors or points represented by $D(M)$ (i.e. input data matrix) and $D(M')$ (i.e. MDS map) is calculated by a stress function. The general form of this function is given by the following equation:

$$stress = \sqrt{\frac{\sum_{i,j=1..m} (\delta_{i,j} - d_{i,j})^2}{\sum_{i,j=1..m} d_{i,j}^2}}$$

As a general rule, the smaller the stress, the better the visual representation. So if the stress is zero, the indication would be that the resulting MDS map represents the original proximity matrix exactly, but this is rarely, if ever, the case; the aim is to minimize the stress function value for the selected threshold k . By iterating steps (3) and (4) above MDS, the value of the stress function is gradually minimized until there is no further reduction and, at which point, the iteration stops.

As with other dimensionality reduction methods, a threshold dimensionality k must be determined for MDS. The sign that k is too small is stress far from 0; stress typically increases as the number of dimensions decreases and vice versa; a 2-dimensional representation usually has more stress than a 3-dimensional one. If $k = n$, that is, the selected dimensionality is the same as the original data dimensionality, the stress will be at or very close to 0. For dimensionality reduction the question is: what should the dimensionality be to give an adequate stress rank? For clustering, as with PCA, this is not an issue, since k must be three or less. The stress value at this dimensionality is a sign of how well the reduced matrix represents the original one, and thereby of how reliable the clustering is likely to be: the higher the stress, the more likely it is that the clustering is based on a poor representation of the original data.

For more on MDS see, for example, Moisl (2015), Borg & Groenen (2005), Jackson (2003), Jolliffe (2002), Kruskal & Wish (1978). For briefer accounts see, for example, Martinez, Martinez, and Solka (2011), Hair et al. (2010), Izenman (2008), Lee & Verleysen (2007), Groenen & Velden (2005), Wickelmaier, 2003, Jain & Dubes (1988).

iii. Isomap:

Isomap is an alternative of MDS (Moisl, 2015; Lee & Verleysen, 2007; Tenenbaum et al., 2000) which 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 (Lee & Verleysen, 2007) is a widely used paradigm in data analysis to approximate geodesic

distance between different points of a manifold using graph distance (Moisl, 2015; Lee & Verleysen 2007).

Mathematically, geodesic distance is a generalization of linear to nonlinear distance measurement in a data space: the geodesic distance $g(x,y)$ is the shortest distance between two points x and y on a manifold measured along its possibly-curved surface (Deza & Deza, 2009). This can be shown in figure (2.29), taken from Moisl (2015:42):

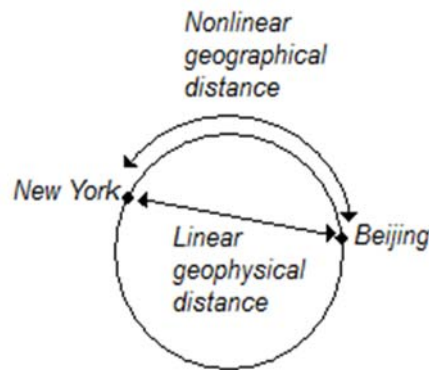


Figure (2.29): Linear geophysical and nonlinear geographical distance between points on the Earth's surface

The Isomap approximation employs the topological concept of neighbourhood. The concept of topology is central to understanding of manifolds in general and of Isomap in particular. It comes from pure mathematics concerned with general properties of metric spaces. Topology studies manifolds as topological spaces and thus defines them as spaces on their own irrespective of any embedding metric space and related axes (Moisl, 2015; Munkres, 2000; Mendelson 1975).

Specifically, topology describes manifold points situated or populated in the metric space of Figure (2.30a) independently both of the metric defined on the space and of the coordinates relative to which the distances among vector points are calculated, as in figure (2.30b).

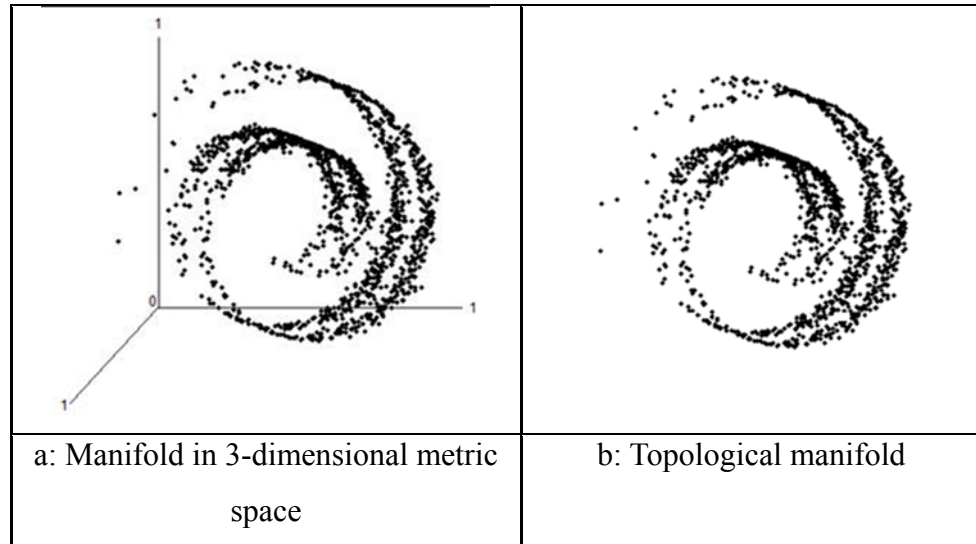


Figure (2.30) A manifold embedded in metric space (a) and as topological manifold (b), taken from Moisl (2015: 125)

Topology changes the concept of metric and related Cartesian coordinates with relative closeness of vector points to one another in the manifold as the mathematical pattern assigned to the underlying set of data points; relative closeness of vector points to each other is defined by a function which, for any given vector point p in the manifold, returns the set of all vector points within some defined proximity to p .

The question to be asked now is, in the absence of a metric and a Cartesian coordinate system, how is the proximity described? The answer is that topological spaces are generated from metric ones and acquire from the latter the concept of neighbourhoods or the notion of closeness. In terms of metric and topological spaces, a subset of vector points which from a topological point of view creates manifold points can itself be divided into subsets of a fixed size called neighbourhoods, where the neighbourhood of a point p in the manifold can be understood either as the set of all vector points within some fixed radius ϵ from p or as the k nearest neighbours of p using the existing metric and coordinates; in figure (2.31) small region of the manifold points from figure (2.30) is zoomed in to show these two types of neighbourhood.

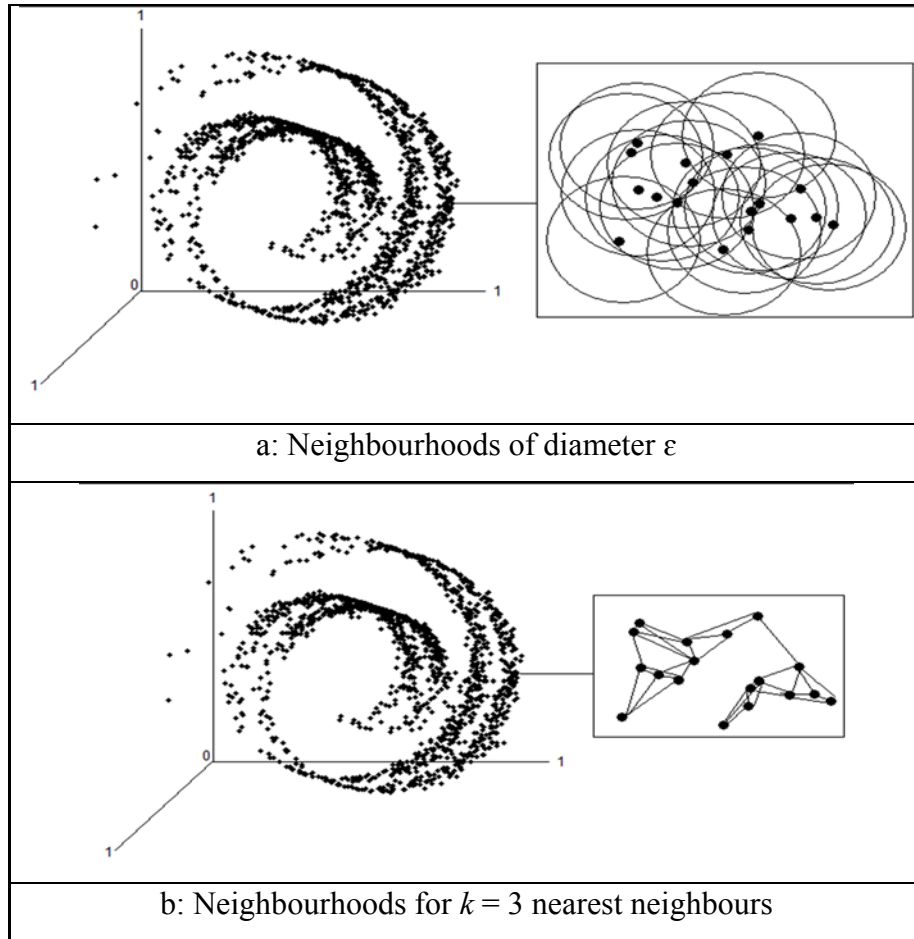


Figure (2.31) Neighbourhoods in a zoomed-in fragment of a geometric object in metric space, taken from Moisl (2015: 125)

In Figure (2.31a), the neighbourhood points are shown as circles within the zoomed-in rectangle where the neighbourhood of every vector point is the other vector points within a radius of ϵ ; in (3.48b), where a neighbourhood of any vector point is the k nearest vector points regardless of distance, the neighbourhood points are shown as lines, for $k = 3$, linking each vector point to the three closest nearest to itself. Once a manifold of points has been grouped or divided into neighbourhoods and thereby converted into a topological space, the frame of reference is ignored and only the neighbourhoods specified in terms of the metric are maintained. In such a manner, point manifolds of arbitrary shape can be understood as being consisted of metric subspaces; if the original metric is Euclidean, for example, the manifold points in figure (2.31) can be seen as flat shapes like a patch work of locally-Euclidean subspaces. It is, therefore, intuitively possible to consider the curved surface of the Earth as similar to a patchwork of flat neighbourhoods as most people see it (Moisl, 2015).

Topological spaces are supersets of metric spaces, so that every metric space is also a topological one. This assumption is taken to make the reference to geometrical objects in subsequent discussion easier and more convenient in which topological spaces are referred to as manifold points regardless of whether they are embedded in a metric space or create a topological space without reference to a Cartesian coordinate system.

To describe how Isomap works based on the concept of topological neighbourhood, we consider the example used in Moisl (2015: 126-33) which shows only one type of neighbourhood, i.e. the k - nearest neighbour. Given an $m \times n$ data manifold M embedded in a metric space and a specification of neighbourhood size as a radius ϵ or as k nearest neighbours, Isomap first converts M into a topological manifold of points by constructing a set of k -neighbourhoods. This can be performed in two stages:

1. We generate a matrix of linear distances between row vectors, that is, we calculate the rows of M ; we suppose that the measure is Euclidean and we call the generated matrix D .
2. We calculate a neighbourhood matrix N based on D , this shows the distance of each of the row vectors M_i ($i = 1..m$) to its k nearest neighbours.

This can be served as an example using the small randomly generated two-dimensional matrix M whose scatterplot shown with row labels in figure (2.32).

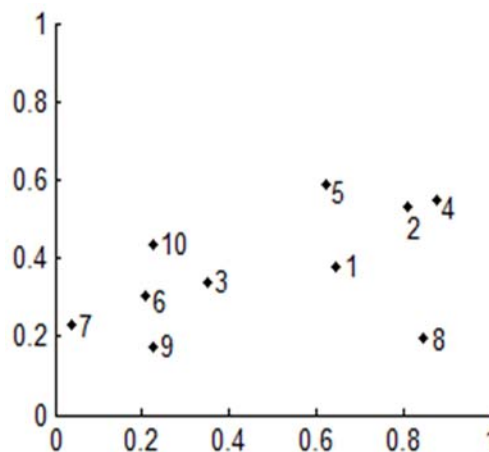


Figure (2.32) Scatter plot of a randomly generated two-dimensional matrix M

Table (2.6a) shows the data matrix M underlying figure (2.32), table (2.6b) the Euclidean distance matrix D for M , and (2.6c) the corresponding neighbourhood matrix N for $k = 4$.

a: M		v1	v2								
	1	0.64	0.37								
	2	0.81	0.53								
	3	0.35	0.33								
	4	0.87	0.55								
	5	0.62	0.58								
	6	0.20	0.30								
	7	0.04	0.23								
	8	0.84	0.19								
	9	0.22	0.17								
	10	0.22	0.43								
b: D		1	2	3	4	5	6	7	8	9	10
	1	0	0.228	0.296	0.288	0.210	0.443	0.622	0.272	0.467	0.421
	2	0.228	0	0.500	0.067	0.197	0.647	0.829	0.340	0.689	0.592
	3	0.296	0.500	0	0.566	0.368	0.148	0.329	0.514	0.210	0.157
	4	0.288	0.067	0.566	0	0.256	0.713	0.895	0.357	0.753	0.658
	5	0.210	0.197	0.368	0.256	0	0.504	0.683	0.451	0.575	0.423
	6	0.443	0.647	0.148	0.713	0.504	0	0.182	0.645	0.132	0.136
	7	0.622	0.829	0.329	0.895	0.683	0.182	0	0.805	0.195	0.278
	8	0.272	0.340	0.514	0.357	0.451	0.645	0.805	0	0.619	0.662
	9	0.467	0.689	0.210	0.753	0.575	0.132	0.195	0.619	0	0.265
	10	0.421	0.592	0.157	0.658	0.423	0.136	0.278	0.662	0.265	0
c: N		1	2	3	4	5	6	7	8	9	10
	1	0	0.228	Inf	0.288	0.210	inf	Inf	0.272	inf	Inf
	2	0.228	0	Inf	0.067	0.197	inf	Inf	0.340	inf	Inf
	3	0.296	inf	0	inf	inf	0.148	Inf	Inf	0.210	0.157
	4	0.288	0.067	Inf	0	0.256	inf	Inf	0.357	inf	Inf
	5	0.210	0.197	0.368	0.256	0	inf	Inf	Inf	inf	Inf
	6	inf	inf	0.148	inf	inf	0	0.182	Inf	0.132	0.136
	7	inf	inf	0.329	inf	inf	0.182	0	Inf	0.195	0.278
	8	0.272	0.340	Inf	0.357	0.451	inf	Inf	0	inf	Inf
	9	inf	inf	0.210	inf	inf	0.132	0.195	Inf	0	0.265
	10	inf	inf	0.157	inf	inf	0.136	0.278	Inf	0.265	0

Table (2.6): a. A matrix M underlying figure (2.32), b. Euclidean distance matrix D for data in table (2.6a), c. Neighbourhood matrix N corresponding to Euclidean distance matrix in table (2.6b), taken from Moisl (2015: 127-128)

Based on the data and distance previous discussions M and D are easy to understand without explanation. N is unobvious and needs some explanation. The first thing we must note is that, apart from 0 in the main diagonal, each row of N has exactly 4 values, which are equivalent to $k = 4$. The value at N_{ij} means both that j is in the k -neighbourhood of i and the distance between i and j ; the k -neighbourhood of N_1 , for example, includes N_2 , N_4 , N_5 and N_8 , which can be visually approved by figure (2.32). The zeros mean that a data object is at a nil distance from itself, and the *inf* values (for 'infinity') that j is not in the neighbourhood of i .

In the framework of interpreting data analysed by Isomap, this method interprets neighbourhood matrix N as a graph in which data vectors are nodes, the values are arcs labelled with distances between pairs of nodes, and the *inf* values mean that there is no arc. In graph representation, the N of table (2.6c) looks like figure (2.33).

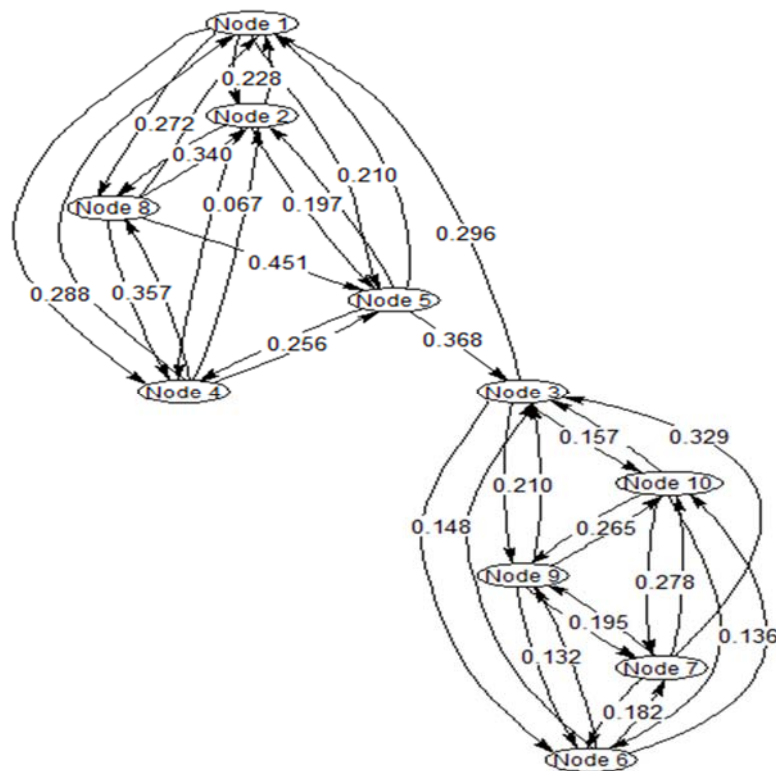


Figure (2.33) Graph interpretation of the neighbourhood matrix in table (2.6c), taken from Moisl (2015: 129)

In this graph, the length of the shortest path that links two points or the shortest node-to-node distance between each pair of points on the manifold can be calculated using one of the standard graph “traversal algorithms”, which are used to examine each node in a resulting tree pattern and check its value (Gross & Yellen, 2006). So, the shortest distance

between node 8 and node 7, for example, follows the path $8 > 5 > 3 > 6 > 7$, and is $0.451 + 0.368 + 0.148 + 0.182 = 1.149$. Referring to table (2.6b) above, it shows that this value is bigger than the Euclidean distance of 0.805, which goes directly from column 8 to row 7. Isomap calculates the graph distances between all combinations of vector points on the manifold and saves them in a matrix G . The one generated from figure (2.33) above is shown in table (2.7) below. It must be noted that the values shown in that table may not be exactly identical with those attainable from figure (2.33) on account of round-off discrepancies. Where there is only a single arc traversal the graph and Euclidean distances are similar but for multi-arc traversals the graph distances are greater where the path between data objects is not linear. Isomap uses the classical or metric least squares MDS procedure to such graph distance matrices G to reduce their dimensionality, as already described.

	1	2	3	4	5	6	7	8	9	10
1	0	0.228	0.578	0.288	0.210	0.725	0.907	0.272	0.787	0.734
2	0.228	0	0.565	0.067	0.197	0.713	0.895	0.340	0.774	0.721
3	0.296	0.524	0	0.584	0.506	0.148	0.330	0.568	0.210	0.157
4	0.288	0.067	0.624	0	0.256	0.772	0.954	0.357	0.834	0.781
5	0.210	0.197	0.368	0.256	0	0.516	0.698	0.481	0.577	0.524
6	0.444	0.672	0.148	0.732	0.654	0	0.182	0.716	0.132	0.136
7	0.625	0.853	0.329	0.914	0.835	0.182	0	0.897	0.195	0.278
8	0.272	0.340	0.819	0.357	0.451	0.966	1.149	0	1.028	0.975
9	0.506	0.733	0.210	0.794	0.715	0.132	0.195	0.777	0	0.265
10	0.453	0.680	0.157	0.741	0.662	0.136	0.278	0.724	0.265	0

Table (2.7): Shortest-path graph distance table for table (2.6c) / Figure (2.33), taken from Moisl (2015:130)

The choice of a dimensionality k and assessment of how well the original distances have been preserved in the reduced-dimensionality representation are similar to that of MDS, and are not repeated here; however, where the classical MDS procedure is used, the criterion for selection of k is residual variance rather than stress. (Moisl, 2015)

It remains, finally, to say that Isomap was proposed by Tenenbaum et al. (2000), and modified to deal with a greater range of nonlinear manifold types in de Silva & Tenenbaim (2003). Other useful accounts are in Moisl (2015), Xu & Wunsch (2009), Lee

& Verleysen (2007).

iv. Self-Organizing Map:

The 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 two-dimensional. 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 (Moisl, 2015; Chattopadhyay et al., 2011; Kohonen, 2001; Hollmen, 1996).

The standard reference work for SOMs is Kohonen (2001). Briefer accounts can be found in Moisl, (2015); Chattopadhyay et al. (2011) and in most artificial neural network textbooks, for example, Silva (2008); Pang (2003); Allinson et al. (2001); Germano (1999); Haykin (1999), see also the papers by Oja and Kaski (1999); Verleysen (1997); Gurney (1997); Mehotra et al. (1997); Kaski (1997); Ritter et al. (1992). What follows is based in large part on these sources, and in particular Moisl (2015).

A SOM consists of three components that are part of it: an input buffer, a two-dimensional lattice of processing units, and connections between the buffer and the lattice, as shown in figure (2.34) below.

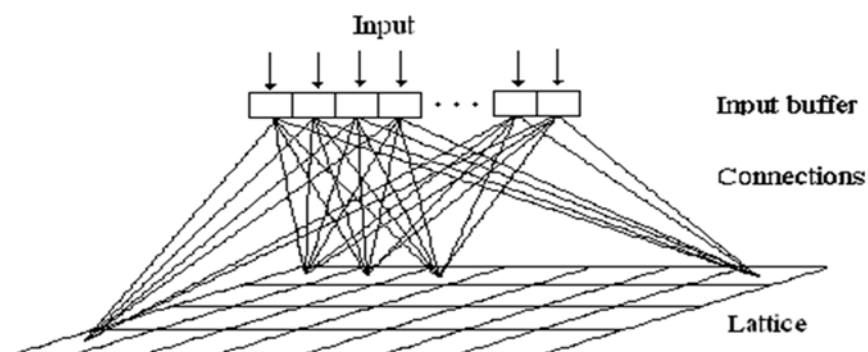


Figure (2.34) Structure of a self-organizing map, taken from Moisl (2015:162)

- The input buffer is a vector v whose length equals the number of empty spaces in the buffer: a buffer with k empty space appears in the mathematical model as a vector $v = [v_1, v_2 \dots v_k]$, where k is a positive integer, and each of the vector elements v_i contains a number that represents the vector component in the corresponding buffer empty space. Figure (2.35) defines a 6-vector, where each components is represented by a real-valued number in the range 0..1:

Vector	.38	.92	.64	.38	.92	.64
	1	2	3	4	5	6

Figure (2.35) SOM input buffer

- The lattice is a 2-dimensional surface of cells represented as a matrix M whose row and column dimensions are the same as those of the lattice, and whose elements contain numbers that represent degree of activation. Figure (2. 36) shows a 4 x 4 lattice, where each cell contains a vector of weights of the same dimension as the input vector and the degrees of activation are represented by real-valued numbers:

Matrix			
.21	.21	.21	.21
.78	.78	.78	.21
.78	.95	.78	.21
.78	.78	.78	.21
1	2	3	4

Figure (2.36) SOM lattice

Particular elements in M are indexed by row and column coordinates i and j with row 1 at the top and column 1 leftmost, as shown; the matrix element with the highest numerical activation value is (.95) and is indexed by $M_{3,2}$.

- The connections are links from the input buffer to the lattice each of which has a particular strength. These connection strengths are fundamental to the operation of the SOM, and are learned from iterative exposure to input vectors via buffer rather than explicitly specified. Relative to an $m \times n$ data matrix D , the learning procedure is as follows:

1. Select a row vector D_i (for $i = 1..m$) and present it to to the input buffer.

2. Propagate the input along the connections to selectively activate the cells of the lattice, where the activation of a given cell is the sum of all the components arriving via the connections converging on that unit.
3. Search the lattice to identify the cell with the greatest activation.
4. Strengthen all the connections converging on the most-activated unit as well all those in its immediate vicinity. The input vector is thereby more strongly associated with the region of the lattice containing the most-activated unit.
5. Repeat (1)-(4) until the connections no longer require strengthening, which indicates that the input data has been learned in the sense that each row vector from the data matrix D has been assigned to a particular region of the lattice.

The above steps give an intuitive account of the SOM learning procedure. Its details are considerably more complex, and can be found in the references given above.

Once the data has been learned, the SOM can be used for clustering. The row vectors from D are again presented in succession, this time without adjustment of the connections, and each activates the specific region of the lattice which the learning procedure has assigned it. After all, the input vectors have been presented, there is a pattern of activations on the lattice; this pattern is the cluster structure of the data. For example, assume that a SOM has been trained using data consisting of 20 input vectors. The clustering stage would generate an activation pattern on the lattice something like that shown in Figure (2.37).

		v11	v12	v5		
		v8	v13	v18		
		v16	v15	v20		
		v3	v9	v7		
		v4	v17	v2		
v19	v14					
v16	v1					
						v10

Figure (2.37) An example of SOM trained on 20 vectors

There are two plausible clusters on the lattice representing the neighbourhood relationships among 20 vectors: the first cluster comprises (v11,v12,v5, v8, v13, v18, v16, v15, v20, v3, v9, v7, v8, v17, v2) and the other comprises (v19, v14, v16, v1). The vectors inside each cluster fall within some spatial adjacency distance from each other; they are differentiated as clusters because they are near each other in the space, or topologically adjacent in the input space and therefore are mapped to nearby map units in the SOM forming clusters. (v10) did not form a cluster with any other vector because it is topologically distant in the input space and therefore are not kept close to other vectors on the map.

To sum up, the SOM's representation of high dimensional data in a low-dimensional space is a two-step process. The SOM is first trained using the vectors comprising the given data. Once training is complete all the data vectors are input once again in succession, this time without training. The aim now is not to train but to generate the two-dimensional representation of the data on the lattice. Each successive input vector activates the unit in the lattice with which training has associated it together with neighbouring units. When all the vectors have been input, there is a pattern of activations on the lattice, and the lattice is the representation of the input manifold in two-dimensional space.

v. Hierarchical clustering:

Hierarchical clustering has been and continues to be the most widely used of the available clustering methods, and so is covered in most accounts of cluster analysis, multivariate analysis, and related disciplines like data mining and information retrieval. A selection of references is (Moisl, 2015; Everitt et al., 2011; Mirkin 2011; Xu and Wunsch 2009; Gan et al., 2007; Tan et al., 2006; Gore, 2000; Jain et al., 1999; Gordon, 1999; Jain and Dubes 1988; Romesburg, 1984). For less depth discussions see, for example, Everitt and Dunn (2001), Gore (2000), Jain et al. (1999), Hair et al. (1998), and Oakes (1998).

The clustering methods (i)-(iv) described thus far have all represented clusters as concentrations of points on a two-dimensional surface and have relied on the innate human pattern perception capability to identify the concentrations as clusters. Hierarchical analysis differs from these in that it represents the distance relations among m objects in an n -dimensional data space as a recursively embedded constituency

structure, that is, as a binary tree or 'dendrogram'. A hierarchical cluster tree for 10 data objects, for example, might look like the one in Figure (2.38).

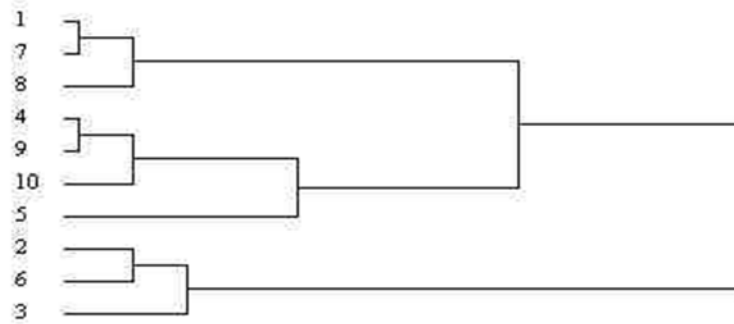


Figure (2.38) Hierarchical clustering tree or dendrogram

In Figure (2.38), the data items are at the leaves of the tree, and the lengths of the horizontal lines joining any two items or subtrees represent the distance between them in the data space. Items (1) and (7), for example, are joined by relatively short lines, indicating that they are close to one another; item (8) is also close to (1) and (7) but not as close as (1) and (7) are to one another, and so is joined to the (1,7) subtree by a slightly longer line; subtrees ((1,7),8) and (((4,9),10),5) are relatively distant from one another and to are joined by relatively long lines; and so on. Such a tree provides an exhaustive representation of the distance relations among data items in a data space. It is up to the analyst to decide where the clusters are; in the above example, the intuitively obvious interpretation is that there are three clusters: the relatively short lines joining the constituents of ((1,7),8), (((4,9),10),5) and ((2,6),3) indicate that these constituents are relatively close to one another in the data space, and the relatively long ones joining the three groups indicate that the groups are relatively distant from one another.

Construction of a hierarchical cluster tree is a two-step process. The first step abstracts a distance table from the data matrix to be analyzed; any distance measure can be used, though Euclidean distance is assumed here. The second step then constructs the tree by successive transformations of the table. The process of transformation is fairly involved and will not be described here; it is discussed in detail in Moisl (2015: 203-8). One aspect of tree construction at the second step does need to be discussed, however: the criterion for joining subtrees. Joining individual data objects is unproblematical— simply join the two closest to one another in the distance table. At subsequent steps in the tree construction process, however, some criterion for judging relative proximity between subtrees is required, and it is not obvious what that criterion should be. Various such

criteria exist, the most often used of which are:

- Single Linkage defines the degree of closeness between any pair of subtrees (X, Y) as the smallest or minimum distance between any of the data points in X and any of the data points in Y.

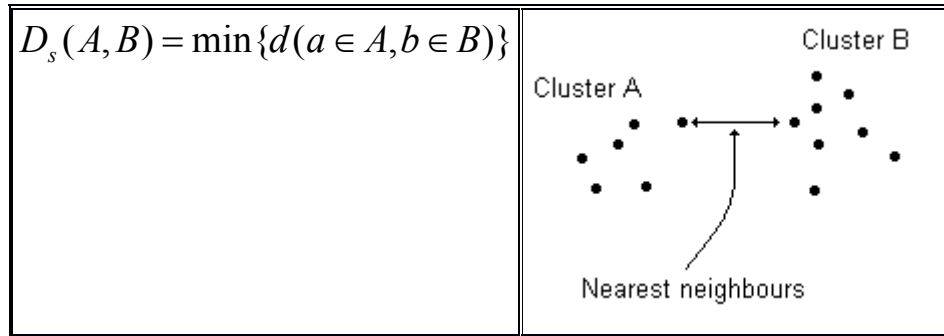


Figure (2.39) Single linkage clustering

- Complete Linkage defines the degree of closeness between any pair of subtrees (X, Y) as the largest or maximum distance between any of the data points in X and any of the data points in Y. The intuition underlying this joining criterion may not be immediately obvious, but it does make sense: finding and joining the cluster pair with the smallest maximum distance between their members creates a cluster with the smallest diameter at that stage in the clustering procedure, and therefore the most compact cluster.

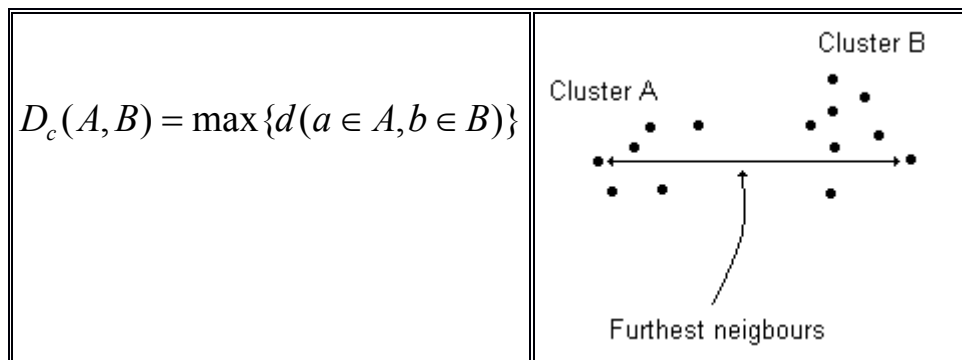


Figure (2.40) complete linkage clustering

- Average Linkage defines the degree of closeness between any pair of subtrees (X, Y) as the mean of the distances between all ordered pairs of objects in X and Y: If X contains x objects and Y contains y objects, the distance is the mean of the sum of (X_i, Y_j) , for $i = 1 \dots x, j = 1 \dots y$.

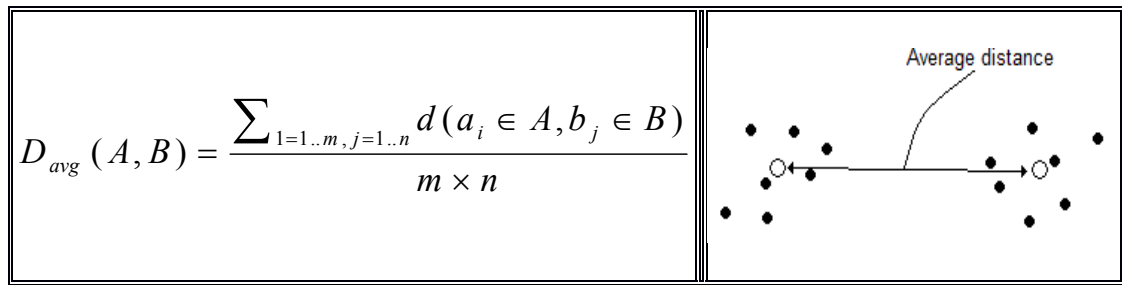


Figure (2.41) average linkage clustering

- Increase in Sum of Squares Linkage (Ward's Method) defines the degree of closeness between any pair of subtrees (X,Y) in terms of minimization of variability relative to an objective function which uses two measures: relative to a cluster A, (i) the error sum of squares (ESS) is the sum of squared deviations of the vectors in A from their centroid, and (ii) the total error sum of squares (TESS) of a set of p clusters is the sum of the ESS of the p clusters. At each step of the tree building sequence, the ESS of the p clusters available for joining at that step is calculated. For each unique combination of cluster pairs the increase in TESS is observed, and the pair which results in the smallest increase in TESS is joined.

Empirical results have repeatedly shown that, relative to any given data matrix, these various joining criteria typically generate trees which differ from one another to greater or lesser degrees. This is to be expected, since each of the criteria is based on a different view of how a cluster should be defined. This raises an obvious question, however: in any given application, which criterion, if any, captures the true cluster structure of the data? This is a fundamentally important question, and it is addressed in the remainder of this section.

What is the 'true cluster structure' of data? This section began by observing that there is no generally agreed formal definition of what a cluster is, and as such this question is itself not well defined. There is, at present, no theoretical basis for identification of true cluster structure. The most that can be said is that, relative to some definition of what a cluster is, any clustering method is more or less successful at identifying clusters in given data. A range of validation methods for assessing the efficacy of the various methods have been and continue to be developed; a recent overview of them is given in Moisl (2015: 224-249). These validation methods have their own problems, however, and not all

researchers accept their effectiveness. The present study therefore takes a different approach to validation of its clustering results: it applies a range of methods to the data, each 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.

In the next chapter the Coleridge data matrix will be analyzed using all these methods, and interpretation will be based on the extent to which the results generated by the different methods agree. Many more methods could, of course, have been used in addition, but inclusion of these would have extended the discussion greatly, and to keep it within reasonable bounds, selection was unavoidable.

Chapter Three

Analysis

This chapter applies the methodology described in the preceding one to test the hypothesis that Coleridge was the author of the 1821 Boosey translation of Goethe's *Faust*. This testing is done in a sequence of steps. The first step creates a corpus of texts by Coleridge from which function word frequency data is abstracted, and the data is then clustered to describe Coleridge's style in the sense that the structure of his usage of function words across the texts in the corpus is established. The second step expands the range of texts to be analysed by including works by Byron, Shelley, and Wordsworth in the corpus; function word frequency data is then abstracted from the expanded corpus and cluster analysed to see how Coleridge's function word usage compares to that of the three contemporary writers selected as comparators, the aim being to see if the concept of a characteristic authorial style which is at the heart of authorship attribution is tenable. The third part adds the 1821 Boosey *Faustus* to the corpus; function word frequency data is then abstracted and clustered to see how *Faustus* fits into the existing cluster structure. The fourth and final step is a close analysis of the immediate neighbours of *Faustus* in the cluster structure. The conclusion is that the hypothesis being tested is falsified.

3.1 Data creation: function word frequency in Coleridge's works:

Digital electronic copies of the texts comprising Coleridge's literary output in prose, verse, and drama were assembled into a corpus. They are 363 raw texts saved in an ASCII (txt.doc) format and will be found in the Appendix (1). However, significant variations in the lengths of these texts were found during the stage of corpus construction. Some texts were large enough in size to be analytically practical. They are 31 texts and will be found in the Appendix (2). Other texts were too short to achieve a good level of analytical accuracy for reasons discussed in the previous chapter section (2.2.3.3/c). They are 332 texts and will be found in the Appendix (3). These texts were amalgamated and assigned into 21 collections of texts according to their appearance in journals and poetry collections. They are treated as unitary texts and will be found in the Appendix (4).

Table (3.1) lists all Coleridge's works considered for the study and shows their electronic sources. For simplicity of exposition, we divided these texts into two main groups: group

(A) includes all the long texts and group (B) includes all the short texts aggregated into 21 collections of poems. For example, *Sibylline Leaves* includes the many short texts poems listed in collection (1), *Juvenile* poems includes the many short poems listed in collection (2), *Adaptations* included the ones listed in collection (3), and so on to the remaining collection of poems. Also, where the name of a given work is long, we referred only to the first word of that work where necessary.

	Name of publication	Works selected	Electronic Source
Group A	N/A (These texts are long texts, each of which is analysed as a text on its own)	Alice 1828, Ancient Mariner 1798, Autumnal 1788, Christabel 1797, Death of Chatterton 1790, Dejection 1802, Delinquent 1824, Departing 1796, Destiny of Nations 1790, Fears 1798, France 1798, Friend 1818, Grenville 1799, Happiness 1791, Improvisatore 1827, Old man 1798, Osorio 1797, Piccolomini 1800, Picture 1802, Pixies 1793, Recantation 1798, Religious Musings 1795, Remorse 1813, Robespierre 1794, Tears 1820, The Nightingale 1798, The Wanderings of Cain 1798, The Three Graves 1798, To Wordsworth 1807, Wallenstein 1800, Zapolya 1816.	Literature Online (Chadwyck) University of Virginia Library
Group B			
1	<i>The Sibylline Leaves: A Collection of Poems by S. T. Coleridge.</i> London: Rest Fenner, 23, Pater-Noster Row. Compositions from	France 1798, The Keepsake 1800, Love 1799, Frost at midnight 1798, Tell's Birth place 1796, Fire, Famine, and Slaughter 1798, To a lady with Falconer's Shipwreck 1814, On receiving the seashore ND, Georgina 1799, A Christmas Carol 1799, Ne plus ultra 1826, To a young lady on her	Literature Online (Chadwyck)

	1793 to 1817, first edition. Curtis printer, Camberwell, London.	recovery from a fever ND, Tranquillity 1801, Human life 1815, Love 1799, Something childish but natural ND, To a young friend on his proposing 1796, The visit of the Gods 1799, The Ballad of The Dark ladie ND, Home sick ND, Observing blossom 1796, Lines to W.L 1797, Elegy imitated from Akenside 1794, Lewti 1798, Answer to a child's question 1820, The Eolian Harp 1795, River Otter ND, Separation 1805, The Night Scene 1813, The Pang more sharp than all 1825, To an unfortunate woman at theatre 1797, To an unfortunate woman 1797, Blank verse Inscriptions 1794, Kubla Khan 1797, Happy husband 1802, Epitaph on an Infant 1794, Limbo 1811, Concert room 1799, Rain 1802, This Lime 1797, Melancholy 1749, The Pains of sleep 1803, The visionary hope 1810, A child's evening prayer 1806, Recollections 1807, Hymn before the sun rise 1802, Lines written in the Hartz Forest 1799, To Rev. George Coleridge 1797, A Tombless epitaph, 1809, To a friend 1794, The Virgin's Cradle Hymn 1811.	University of Virginia Library
2	<i>Juvenile Poems published in The Poetical Works of S. T. Coleridge. By S. T. Coleridge. In Three Volumes. London: William Pickering,</i>	On receiving/hearing account 1791, Lover complaint 1792, Frenzy 1794, Nina thama 1793, Anthem for the children 1789, Gentle look 1793, Easter holiday 1787, Time, real and imaginary 1812, Pain 1790, To the Author of Robbers 1794, Music 1791, Life 1789, Quae verse 1789,	

1840.		<p>Christening a friend's child 1796, Devonshire road 1791, To a Young Ass 1794, Death of Starling 1794, Walk before supper 1792, Inside the coach 1791, The Kiss 1803, Mathematical problem 1791, Evening star 1790, The Sigh 1794, Welsh 1794, The nose 1789, Dura 1787, An infant 1794, Sonnet on quitting school 1791, To the Muse 1789, Amelia 1792, On seeing A youth 1791, The Rose 1793, On a discovery made too late 1794, On Bala Hill 1794, An invocation 1790, A lesson to Englishmen 1795, In the manner of Anacreon 1792, Nil Pejus Est Caelibe Vita 1787, Domestic peace 1794, On an infant 1799, On imitation 1791, Honour 1791, Progress of Vice ND, Lines on a friend who died of a frenzy fever 1794, Blank verse inscription 1794, To disappointment 1792, An effusion at evening 1792, On A lady weeping 1790, On a ruined house 1797, Lines composed while climbing Brockley 1795, Lines in the manner of Spencer 1795, Lines written at Shurton Bars 1795, To a friend in an answer to a melancholy letter 1795, To simplicity 1797, Reflections on Having Left a Place of Retirement 1795, To the Author of poems 1795, Monody on a Tea-kettle 1790, Destruction of Bastille 1798, Ossian 1793, To the Nightingale 1798.</p>	
Poems first published, or re-published, in newspapers or periodicals			

3	<i>Adaptations (1818-1834)</i>	Fulke Greville Lord Brook 1810, On the immortality of soul ND, Letter to Henry ND, The poetaster 1796, Epistle to Sir Thomas Egerton, Knight 1816, On Unworthy wisdom ND, Prologue 1794, Translation of Wrangham's Hendecasyllabi 1794.	Literature Online (Chadwyck) University of Virginia Library
4	<i>Literary Remains (1818-1834)</i>	Julia 1789, To the Rev. W J Hort 1795, Letter to Joseph Cottle 1814, The Rash conjurer 1814, Translation of Ottfried's metrical of the Gospel, I yet remain 1793, Pity 1795, Morienti Suerstes 1798, Psyche 1808, Israel's Lament 1817, Sentimental ND, Inscription for A Time Piece, Epitaphium Testamentarium 1826.	Literature Online (Chadwyck) University of Virginia Library
5	<i>Early Recollections (1837)</i>	To A friend who had declared his intention 1796, The Silver Thimble 1795, From the German 1799.	Literature Online (Chadwyck) University of Virginia Library
6	<i>The Watchman (1796)</i>	To A young lady on her recovery from a fever 1794, Ad Lyram 1794, The hour when we shall meet again 1795, Ode 1792, Lines to a beautiful Spring in a village 1794.	Literature Online (Chadwyck) University of Virginia Library
7	<i>The Cambridge Intelligencer (1794-1798)</i>	Absence- A Farewell 1791, Anna and Harland 1790, Maid of my Love, Sweet Genevieve! 1790, Addressed to a young man of a fortune 1796, Parliamentary Oscillators 1798, Lines written at the Kings	Literature Online (Chadwyck) University of Virginia Library

		arms 1794.	
8	<i>The Morning Post</i> (1797-1800)	The Raven ND, The Devil's thoughts 1799, The two round spaces 1800, To Lesbia 1800, The Mad monk 1800, The Day dream 1802, Moriens Superstiti 1794, Inscription for a seat by the road side 1800, A Stranger Minstrel 1800.	Literature Online (Chadwyck) University of Virginia Library
9	<i>The Courier</i> (The Friend 1809, The Gentleman's magazine 1815, Felix Farley's Bristol Journal 1818, Co-operative magazine and Monthly Herald 1826-1827) (1804-1831).	The Exchange 1804, Pantisocracy 1794, Farewell to love 1805, Pantisocracy in America 1794, The Hour-glass 1811, Fancy in Nubibus 1817, Mutual Passion ND, Apologia pro Vita Sua 1800.	Literature Online (Chadwyck) University of Virginia Library
10	<i>The Morning Chronicle</i> (1793-1795)	On buying a Ticket in the Irish Lottery 1793, epitaph on an infant 1794, Characters (LA FAYETTE) 1794, To the honorable Mr. ERSKINE 1794, To Burke 1794, Priestley 1794, On Pitt and Fox 1806, To the Rev.W. L. Bowles 1794, Siddons 1794, Letter to William Sotheby 1828, To Richard Brinsley Sheridan 1795, To Earl Stanhope 1795.	Literature Online (Chadwyck) University of Virginia Library
11	<i>The literary Souvenir</i> (1826-1829)	Lines suggested by the last words of Berengarius 1826, Youth and age 1823, A day-dream 1802, The two Founts 1826, What is Life 1805, Love's Burial-place	Literature Online (Chadwyck) University of Virginia Library

		1828, Work without hope 1825.	
12	<i>The Friendship's Offering (1834) and Literary Magnet (1827)</i>	My Baptismal birthday 1833, Hymn to the earth 1799, Hexameters 1798, Lines written to Miss Barbour 1829, The Garden of Boccaccio 1828, The Nativity 1827, Hexameters 1798, Water ballad 1799, The Reproof and reply 1823, Sancti Pallium Dominic 1826, Lines to a comic author 1825, Song of a lady's beauty 1830, The faded Flower 1794, An allegoric Romance 1833.	Literature Online (Chadwyck) University of Virginia Library
13	<i>The Anthology published by Thomas Rowley in 1794 and The Anthology published by Francis Wrangham 1795</i>	Monody on the death of Chatterton 1790, To Miss Brunton 1794.	Literature Online (Chadwyck) University of Virginia Library
14	<i>The An Old Man's Diary by Payne Collier, 1871, 2 and Early Recollections by Joseph Cottle, 1837</i>	A character 1825, The knight's tomb 1817.	Literature Online (Chadwyck) University of Virginia Library
15	<i>Epigrams and Jeux' Despart. Taken from The Complete Poetical Works of S. T. Coleridge, including poems and versions of poems</i>	Epigram 59 ND, Epigram 64 ND, Epigram 73 ND Epigram 68 ND, Epigram 1806 (The taste of times), to be sung by the lovers 1801, Drinking vs thinking 1801, The wills of the wisp ND, From an old German poet 1802, on the curious circumstance ND, To my candle 1802,	Literature Online (Chadwyck) University of Virginia Library

	<p><i>now published for the first time in two volumes. Vol. I Poems. Vol. II Dramatic works and appendices</i> Edited by Ernest Hartley Coleridge. Oxford, At the Clarendon Press, 1912.</p>	<p>Epigram on the Secrecy ND, To a lady who requested me to write a poem upon nothing 1822, Authors and publishers 1825, Ideas 1830, Epitaph on himself 1803, Modern critics ND, Written in an Album ND, My God mother's Beard 1791, an invitation to Pool 1797, To a well-known musical critic ND, To captain Findlay 1804, To Susan Steele 1829, Cholera cured before-hand ND, The alternative 1825, On Donne's poetry 1818.</p>	
16	<p><i>Miscellaneous and Later poetry.</i> Taken from <i>The Poetical Works S. T. Coleridge; Reprinted from The Early Editions with Memoir, Notes, etc.</i> London: Frederick Warne and Co. and New York. The presumed publication data of this edition is 1895)</p>	<p>A Lament 1805, Duty surviving Self-Love 1826, Song 1825, Phantom or fact 1830, Constancy to an Ideal Object 1825, The Suicide's Argument 1811, A Soliloquy of the Full Moon 1800, The Madman and 1809, Charity in thought 1830, On my joyful departure 1828, Epilogue ND, First advent of love 1824, Ad Vilmum Axiologum 1805, A Hymn 1814, Forbearance 1832, Motto to 'A Lay Sermon 1817, An angel visitant 1801, An exile 1805, To Asra 1801, On receiving a letter informing me 1796, Coeli Enarrant 1830, Cologne 1828, Desire 1830, Epitaph 1833, Homeless 1826, Humility of the mother of Charity 1830, Self-knowledge 1832, Love and friendship opposite 1830, For a market-clock 1809, To Miss A. T. 1828, Unnamed ND, The outcast 1794, The snow-drop 1800, Faith, hope, and charity 1815, Mahomet 1799, Moto 1808, Not a home 1830, Of human learning stanza</p>	<p>Literature Online (Chadwyck) University of Virginia Library</p>

		1810, Phantom 1805, The presence of love 1807, Epitaph of the present year 1833, Reason 1830, Rossetti ND, Alcaeus to Sappho 1800, The second birth 1801, Sonnet 1805, A sun set 1805, Thomas Hill ND, Thomas Pool 1796, To Marry Pridham 1827, Stanzas ND, A Wish 1792, To the young artist 1833.	
17	<i>Fragments from a note book (1796-1798; 1810-1836).</i> Taken from on <i>The Collected Works of S. T. Coleridge, part I</i> edited by J. C. C. Mays (2001). Princeton University Press.	The night-mare death in life ND, A beck in Winter ND, Not a critic but a judge ND, De Profundis Clamavi 1806, An ode on Napoleon ND, Epigram on Kepler 1799, Translation of the first Strophe 1815, Translation of a fragment of Heraclitus 1822, Imitated from Aristophanes 1816, To Edward Irving 1825, Luther 1826, The Netherlands ND, The Three sorts of friends 1835, A simile ND, Baron Guelph of Adlestan ND, Fragment 3 ND, Fragment 4 ND, Fragment 5 ND, Fragment 6 ND, Fragment 7 ND, Fragment 8 ND, Fragment 9 ND, Fragment 10 ND, Fragment 11 ND, Fragment 12 ND, Fragment 13 ND, Fragment 14 ND, Fragment 15 ND, Fragment 18 ND, Fragment 21 ND, Fragment 1810, Fragment 1792.	Literature Online (Chadwyck) University of Virginia Library
18	<i>Lyrical Ballads</i> (1798 edition)	The Foster-mother's tale 1797, The Dungeon 1796, 1798.	Literature Online (Chadwyck) University of Virginia Library

29	<i>Biographia Literaria</i> (1817)	Biographia 1817, Prose style 1818.	Literature Online (Chadwyck) University of Virginia Library
20	<i>Metrical Feet</i> . Taken from <i>The Collected Works of S. T. Coleridge, Part I</i> edited by J. C. C. Mays (2001). Princeton University Press	A metrical accident 1826, Trochaics 1808, iambic 1801, No sense ND, No sense ND, No sense ND, Plaintive movement 1814, Songs of Shepherds ND, An experiment for metre 1801, Metrical feet Lesson for a boy 1806.	Literature Online (Chadwyck) University of Virginia Library
21	<i>Unfinished letters</i> . Taken from <i>The Collected Works of S. T. Coleridge, Part I</i> edited by J. C. C. Mays (2001). Princeton University Press	Letter to The Rev. H. F. Cary 1818, Letter to James Gillman 1825, Letter to Thomas Poole 1801, Letter to John Thelwall 1796, Letter to C. A. Tulk 1818, To Nature 1820, and verses addressed to J. Horne Took 1796.	Literature Online (Chadwyck) University of Virginia Library

Table (3.1) A selection of Coleridge's works in drama, poetry, and prose

For copyright reasons, none of the publicly-available online digital electronic texts listed in Table (3.1) are based on the most or even relatively recent editions. These texts are taken from *Literature Online (Chadwyck)* and the *University of Virginia Library*:

- <http://lion.chadwyck.co.uk/searchQuickPhase1.doQuickSearchField=coleridge+poetical+works>
- http://xtf.lib.virginia.edu/xtf/view?docId=chadwyck.ep/uvaBook/tei/cheap_3.1452.xml

The origin of these electronic texts is *The Complete Poetical Works of Samuel Taylor Coleridge, including poems and versions of poems now published for the first time edited with textual and bibliographical notes in Two Volumes. Vol. I Poems. Vol II Dramatic Works and Appendices* edited by Ernest Hartley Coleridge and Published in 1912 by The Clarendon Press. Nevertheless, before relying on these electronic texts, it was important to check or examine them for accuracy and make sure that the information or content provided by these texts are free from any corrupted samples (authorial, editorial, and experimental) or any transmission errors occurred by copying or scanning them. For this reason, the online digitized texts were proof-read by carefully comparing them to Ernest Hartley Coleridge's 1912 print edition. This step was necessary to ensure accuracy in our analysis's results because Coleridge, through his writing career which lasted from 1787 to the end of 1832, is known for his textual instability. For each one of Coleridge's poems we have not just a single text but many versions, drafts and alternative versions created by Coleridge himself or by publishers with or without textual authority (Stillinger, 1994). However, the comparison shows that the actual lexical content of the online digital electronic editions doesn't change much from edition to edition, and lexical content is all the researcher is interested in.

As for the *Faustus* translation, the electronic text provided by Oxford University Press 2007 is used, which is available at:

- <http://uk.catalogue.oup.com/product/9780199229680.do>.

For the reasons stated above, the researcher proof-read this electronic text by comparing it to its publically-available printed edition: *Faustus: from the German of Goethe. London: Boosey and Sons, 1821*.

These 21 texts, together with the 31 long texts were now comparable to each other. The next step was to pre-process them prior to constructing the corpus.

The total of 52 digital texts was stripped of textual inclusions not original to Coleridge such as editorial comments and footnotes, line numbers, and so on. This was done computationally using software CLEAN TEXTS shown in the Appendix (11) and the results were subsequently proofread to correct any remaining errors or omissions. A sample an original text and the corresponding cleaned text is given in Table (3.2):

Original text	Corresponding cleaned text
Faust.txt	Clfaust.txt
Ancient mariner.txt	Clancientmariner.txt

Table (3.2): small side-by-side sample of original and corresponding cleaned text

A set of 265 function words to be counted in the Coleridge corpus was then defined. Using the digital Coleridge corpus in conjunction with a digital version of the function words list, a 52 x 193 data matrix D was computationally generated by software called GENERATAE MATRIX shown in the Appendix (12), where each of the 52 rows of D represents a different Coleridgean text, each of the 193 columns represents a different function word, and the value at any $D_{i,j}$ (for $i = 1..52, j = 1..193$) is the number of times that function word j occurs in text i ; the reason that there are only 193 columns in D rather than the full 265 words is that only 193 of the 265 actually occur in the corpus. The generated set of function words is shown in Table (3.3).

Determiners	<i>neither, many, much, various, little, whenever, whatever, whoever, several, both, that, the, their, theirs, these, this, those, wherever, an, all, another, any, enough, each, either, every, few, her, he, hers, herself, him, himself, his, ours, she, my, it, its, itself, me, mine, myself, some, anything, everything, your, our, yours, yourself, yourselves, other, none, they, we, them, themselves, us, something, such, what, which, whom, whose, you, more, less, most, no, certain.</i>
Conjunctions	<i>after, before, behind, below, beneath, beside, besides, and, as, down, during, so, up, upon, of, off, on, since, than, till, until, near, with, within, without, toward, towards, under, underneath, nor, or, though, thus, unless, along, alongside, unto, to, aside, where, whereas, for, from, between, beyond, onto, although, among, amongst, but, by, over, round, around, if, into, except, at, because, whether, while, whilst, since.</i>
Adverbs	<i>however, thence, nevertheless, yet, therefore, when, accordingly, consequently, then, opposite, out, outside, past, nothing, part.</i>
Prepositions	<i>about, above, absent, across, against, amid, amidst, anti, astride, bar, concerning, failing, following, given, including, inside, like, minus,</i>

	<i>respecting, plus, unlike, excluding, save, saving, through, throughout.</i>
Modals	<i>can, could, dare, may, might, must, ought, shall, should, will, would.</i>
Numbers	<i>One, once</i>

Table (3.3) A list of 193 function words

A fragment of D is shown in Table (3.4).

	1 the	2 we	3 of	...	193 whereas
1 Adaptations	55	6	31	...	0
2 Alice	68	0	15	...	0
3 Ancient Mariner	407	17	67	...	0
...
52 Zapolya	799	37	352	...	0

Table (3.4) A fragment of a 52 x 193 data matrix D

D has to be transformed in the two ways described in the *Methodology* chapter prior to analysis.

i. Normalization:

There is a very substantial variation in the lengths of the texts in the Coleridge matrix D. This is shown in Figure (3.1), where the vertical axis represents text length and the horizontal axis the 52 texts arranged in descending order of length.

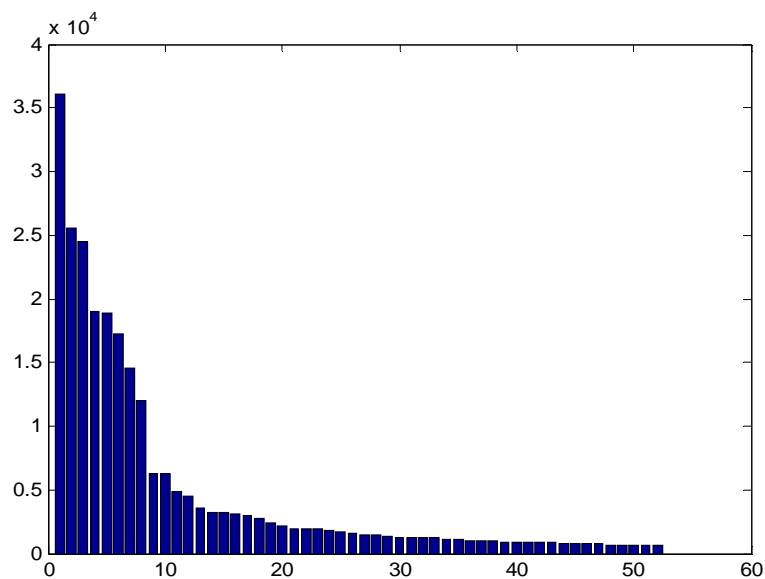


Figure (3.1) Variation in the lengths of the texts in the Coleridge matrix D

This disparity of length, if uncorrected in D , severely skews any clustering results based on D . For example, Figure (3.2) shows a Ward's Method hierarchical analysis of D .

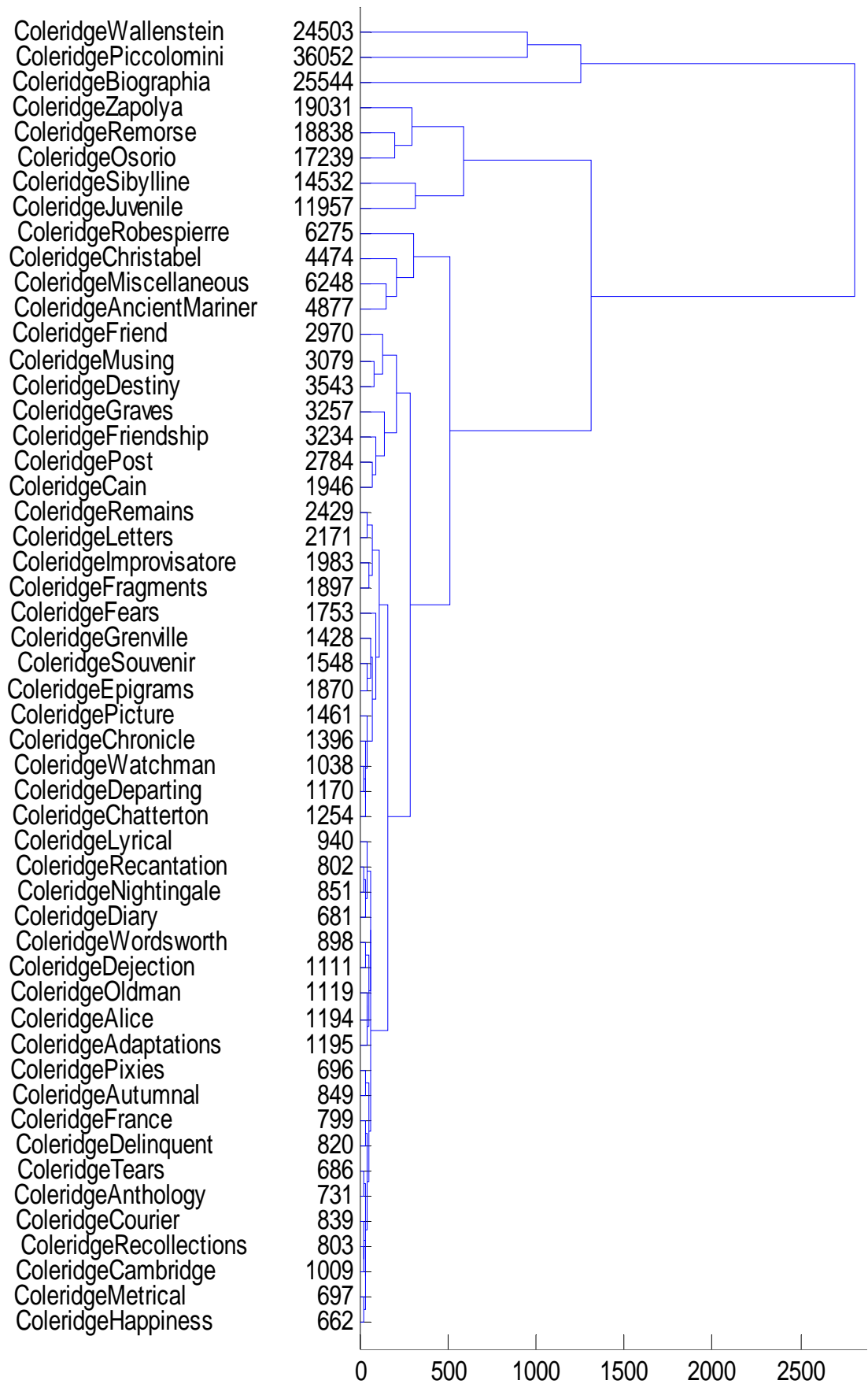


Figure (3.2) Ward's hierarchical analysis of Coleridge's matrix D

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. A programme called EDIT MATRIX, shown in the Appendix (13), was used for the purpose of data normalization.

ii. Dimensionality reduction:

Figure (3.3) shows the distribution of function word frequencies in F1, sorted in descending order, where the vertical axis represents frequency and the horizontal one the column frequencies.

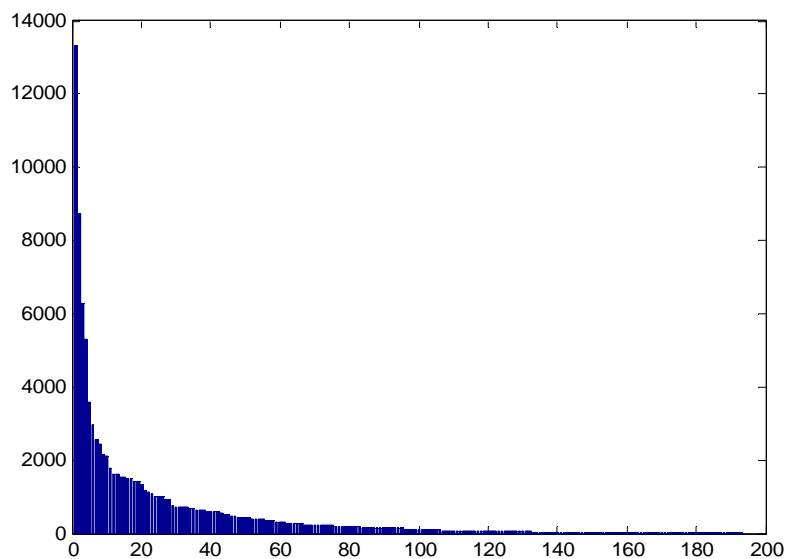


Figure (3.3) The distribution of function word frequency matrix F1

Figure (3.3) shows that there are a few relatively high-frequency function words, a moderate number of medium-frequency ones, and a large number of low-frequency 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. A programme called EDIT MATRIX, shown in the Appendix (13), was used for the purpose of dimensionality reduction.

The order in which these transformations are applied to D is important. If normalization is applied before dimensionality reduction, the normalization procedure would disproportionately increase the values of the low frequency values, as explained in the discussion of normalization in the *Methodology* chapter. This would assign an undue importance to these low-frequency values and would consequently adversely skew the clustering results. Dimensionality reduction is, therefore, applied first to eliminate the

low-frequency columns from D, and normalization is subsequently applied to the dimensionality-reduced matrix. Figure (3.3) above indicates that the 80 most frequent columns is a reasonable choice for retention; the resulting matrix is designated M80Norm to distinguish it from the original matrix D.

3.2 Coleridge's usage of function words:

Having created a data matrix representing Coleridge's usage of function words across his body of work, the first stage of analysis was to determine whether or not there is any discernible structure in that usage. This was done by cluster analyzing M80Norm using the methods outlined in the foregoing chapter. M80Norm was first hierarchically cluster analysed, the results of which are shown in Figures (3.4), (3.5), (3.6), and (3.7). The correspondence of abbreviated labels to full text names is given in Table (3.5).

No.	Text name	Abbreviation
1	Alice 1828	Alice
2	Ancient Mariner 1798	Ancient Mariner
3	Autumnal 1788	Autumnal
4	Christabel 1797	Christabel
5	Death of Chatterton 1790	Chatterton
6	Dejection 1802	Dejection
7	Delinquent 1824	Delinquent
8	Departing 1796	Departing
9	Destiny of Nations 1790	Destiny
10	Fears 1798	Fears
11	France 1798	France
12	Friend 1818	Friend
13	Grenville 1799	Grenville
14	Happiness 1791	Happiness
15	Improvvisatore 1827	Improvvisatore
16	Old man 1798	Oldman
17	Osorio 1797	Osorio
18	Piccolomini 1800	Piccolomini
19	Picture 1802	Picture
20	Pixies 1793	Pixies
21	Recantation 1798	Recantation
22	Religious Musings 1795	Musing

23	Remorse 1813	Remorse
24	Robespierre 1794	Robespierre
25	Tears 1820	Tears
26	The Nightingale 1798	Nightingale
27	The Wanderings of Cain 1798	Cain
28	The Three Graves 1798	Graves
29	To Wordsworth 1807	Wordsworth
30	Wallenstein 1800	Wallenstein
31	Zapolya 1816	Zapolya
32	Juvenile Poems	Juvenile
33	Sibylline Leaves	Sibylline
34	Miscellaneous and Later Poetry	Miscellaneous
35	Fragments	Fragments
36	Epigrams and Jeux D'esprit	Epigrams
37	Literary Remains	Remains
38	Friendship's Offering and Literary Magnet	Friendship
39	Morning Chronicle	Chronicle
40	Metrical Experiments or Feet	Metrical
41	Morning post	Post
42	Cambridge Intelligencer	Cambridge
43	Early Recollections	Recollections
44	Adaptations	Adaptations
45	An Old Man's diary	Diary
46	Literary Souvenir	Souvenir
47	The Watchman	Watchman
48	Lyrical Ballad	Lyrical
49	Anthology	Anthology
50	Biographia	Biographia
51	Letters	Letters
52	The Courier	Courier

Table (3.5) Full names and corresponding abbreviations of Coleridge's texts

From now on, the study used these abbreviations to refer to Coleridge's texts across the various clustering analyses. All the clustering analyses that follow were done by a programme called MATLAB version R2013a, shown in the Appendix (9).

Single Linkage (Cophenetic correlation coefficient: 0.7125):

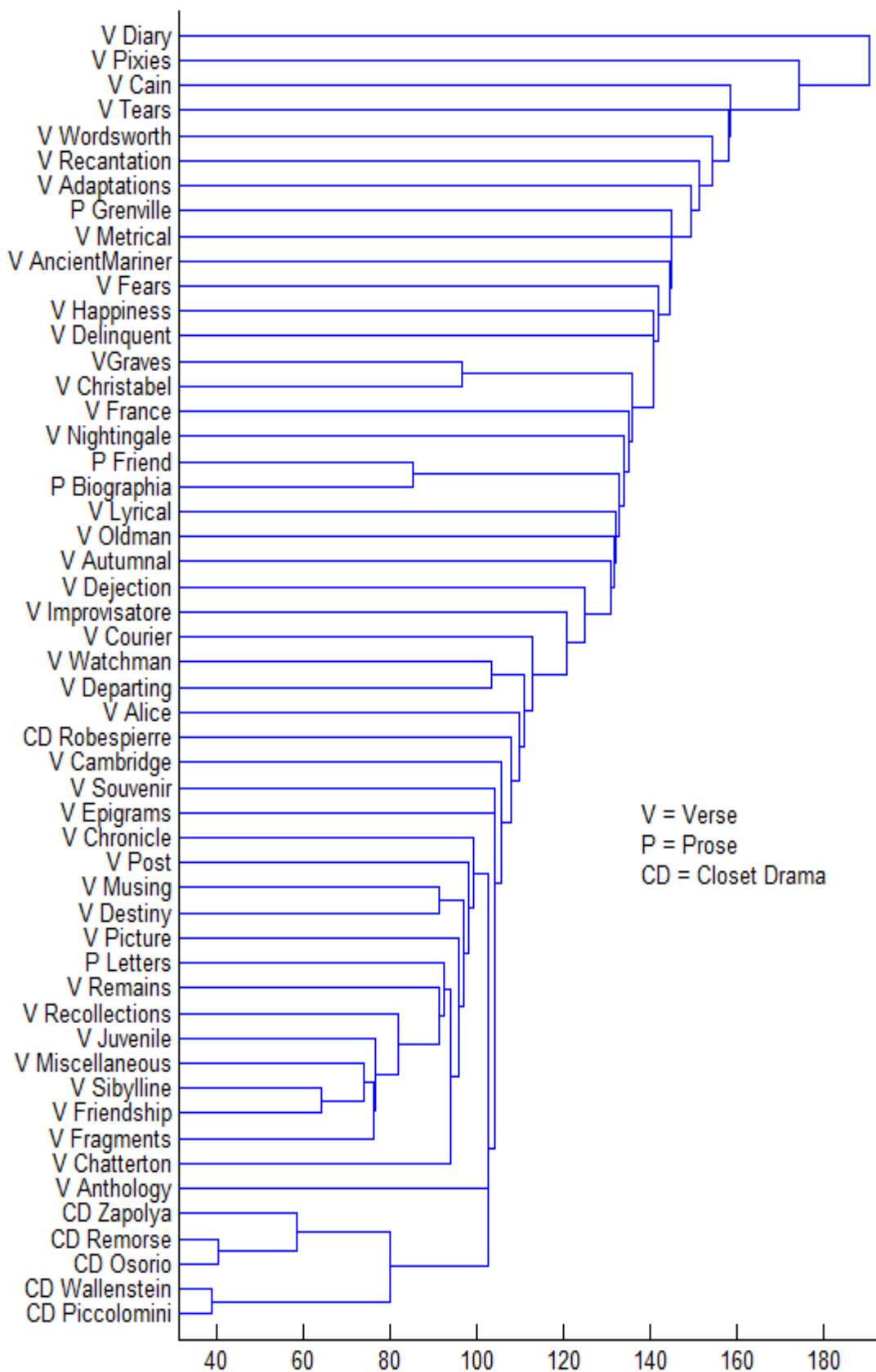


Figure (3.4) Single Linkage. Cophenetic correlation coefficient: 0.7125

Complete Linkage (Cophenetic correlation coefficient: 0.4891):

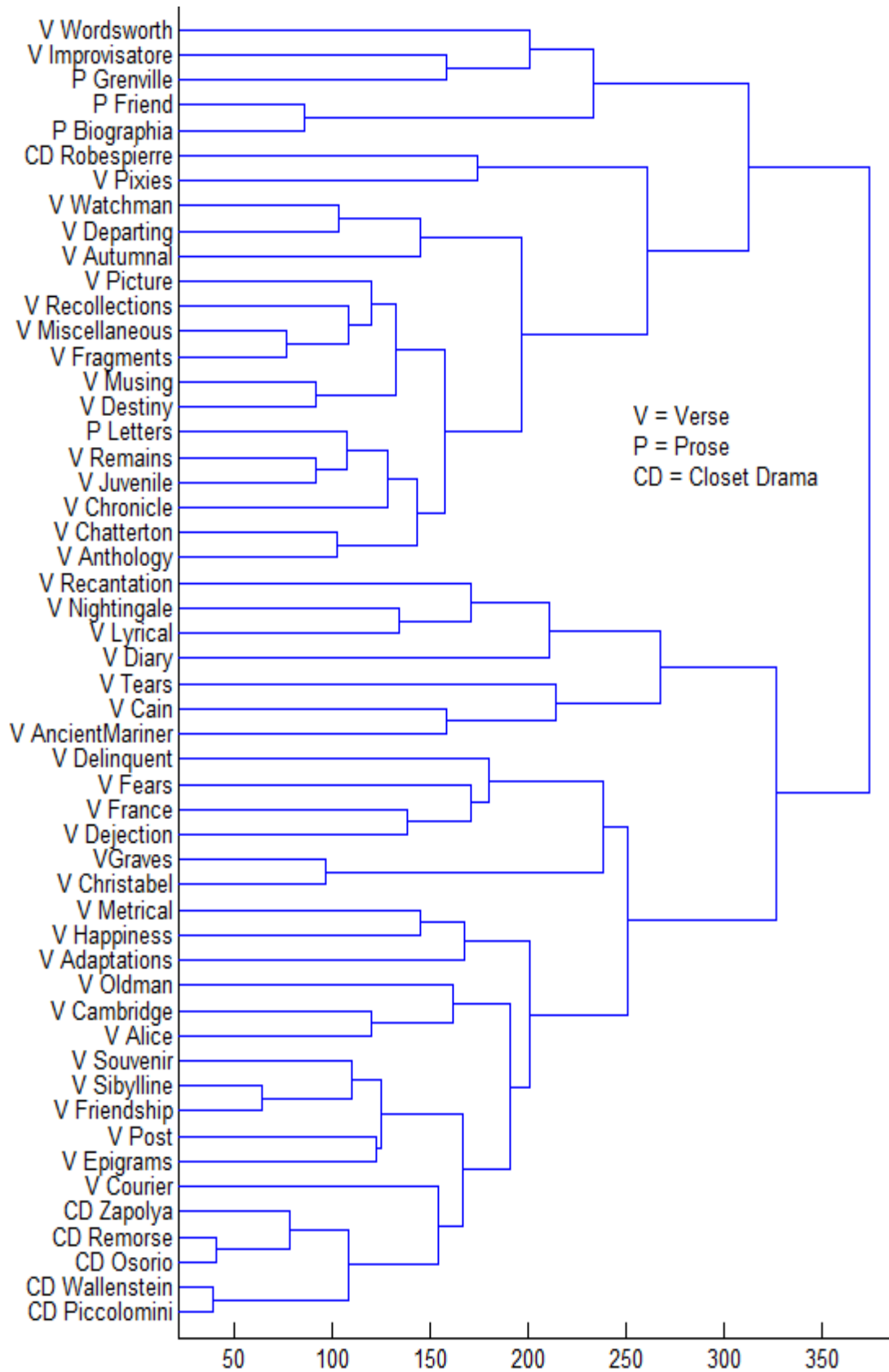


Fig. (3.5) Complete Linkage. Cophenetic correlation coefficient: 0.4891

Average Linkage (Cophenetic correlation coefficient: 0.7694):

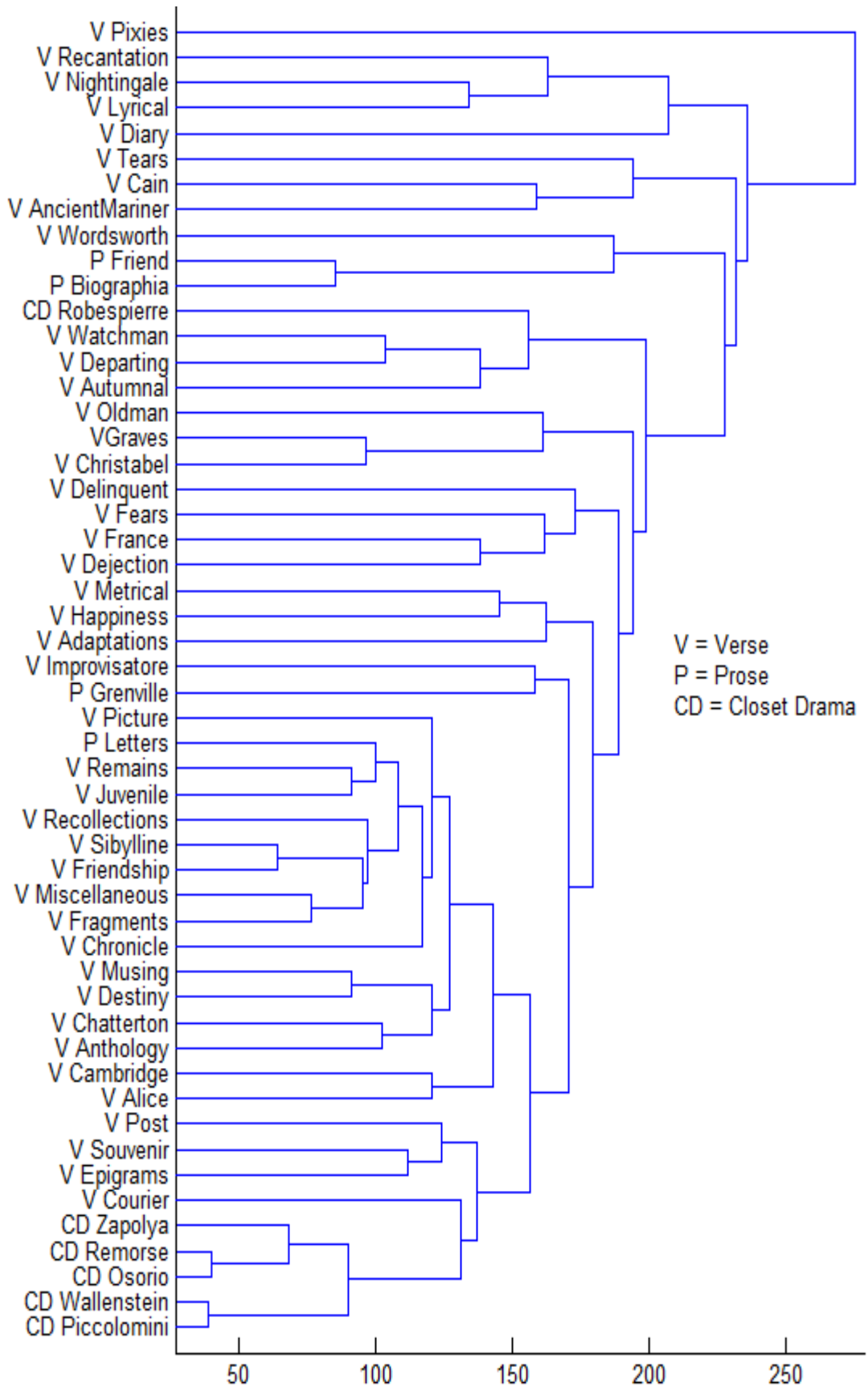


Figure (3.6) Average Linkage. Cophenetic correlation coefficient: 0.7694

Ward linkage (Cophenetic correlation coefficient: 0.5384):

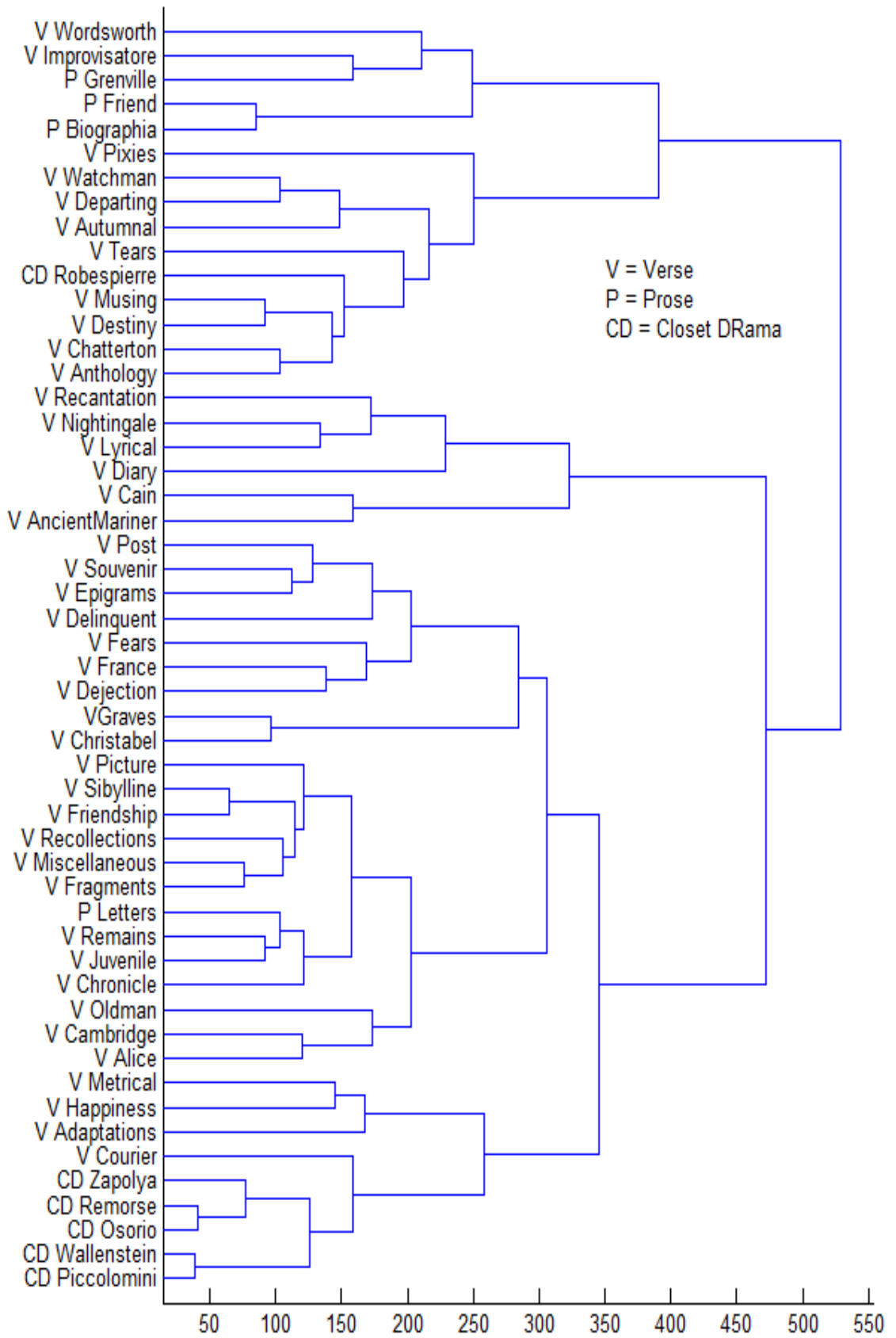


Figure (3.7) Ward linkage. Cophenetic correlation coefficient: 0.5384

The label for each of the foregoing trees includes a value for the associated cophenetic correlation coefficient. This coefficient (Rohlf, 1974; Baker & Hubert, 1974; Sneath & Sokal, 1963; Sokal & Rolf, 1962); summary account in Moisl (2015: 240-4) is one of the cluster validation methods referred to in the foregoing chapter, and is a measure of how well the structure of the tree preserves the distance relations among data objects in the underlying distance matrix. Its range is 0...1, with 1 as perfect preservation; the closer to 1 the coefficient is, therefore, the better the clustering in this sense.

The tree generated by Average Linkage for M180Norm is best for this criterion, though the reservations about the reliability of the cophenetic correlation coefficient noted in Moisl (2015: 240-4) must be kept in mind when assessing the significance of this.

Hierarchical clustering method	Cophenetic correlation coefficient
Single	0.7125
Complete	0.4891
Average	0.7694
Ward	0.5384

Table (3.6) Cophenetic correlation coefficient for of M180Norm and for four hierarchical clustering analyses

Further validation is provided by the range of non-hierarchical clustering methods: PCA, MDS, Isomap, and SOM.

PCA:

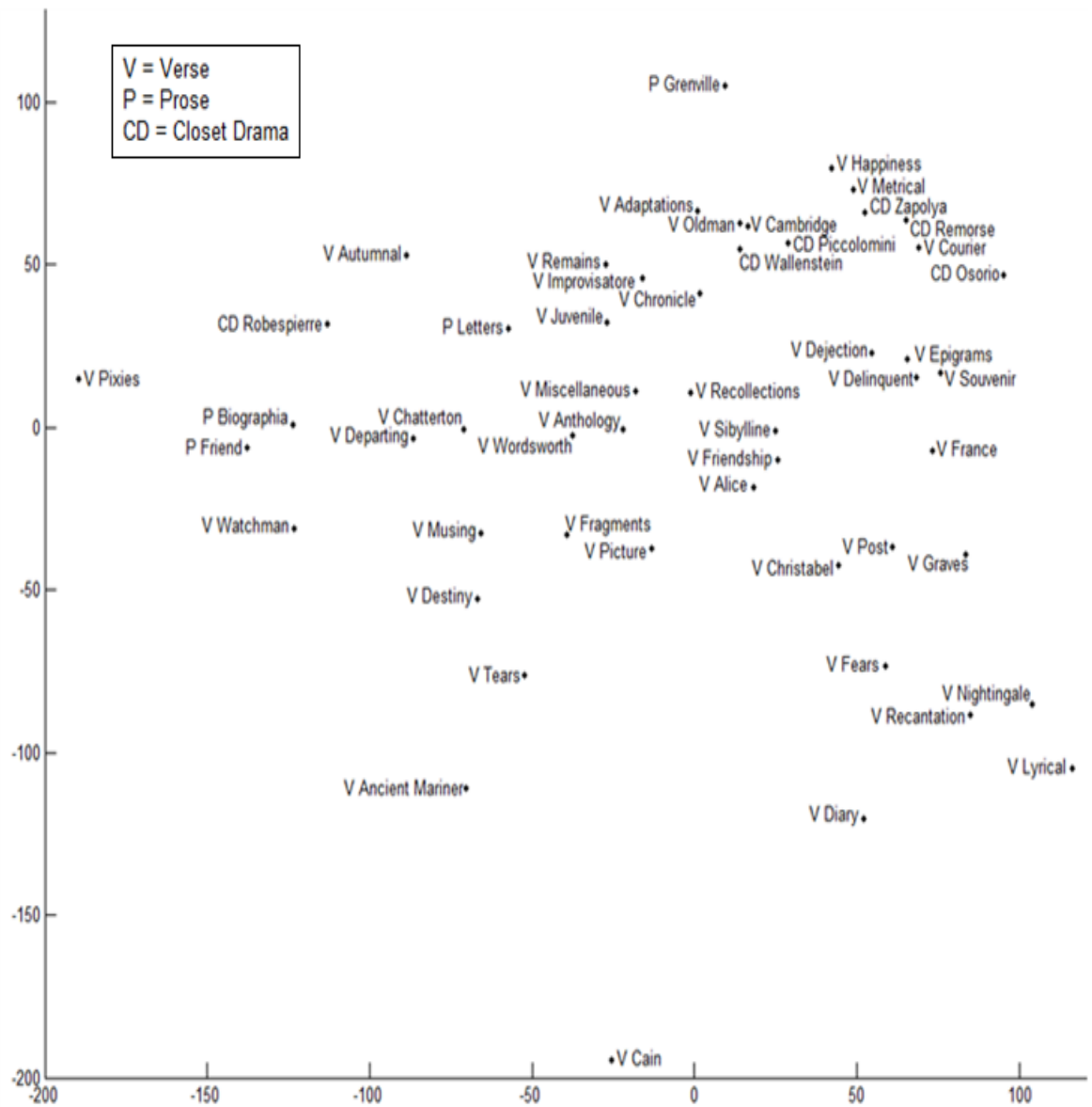


Figure (3.8) PCA of M180Norm

MDS:

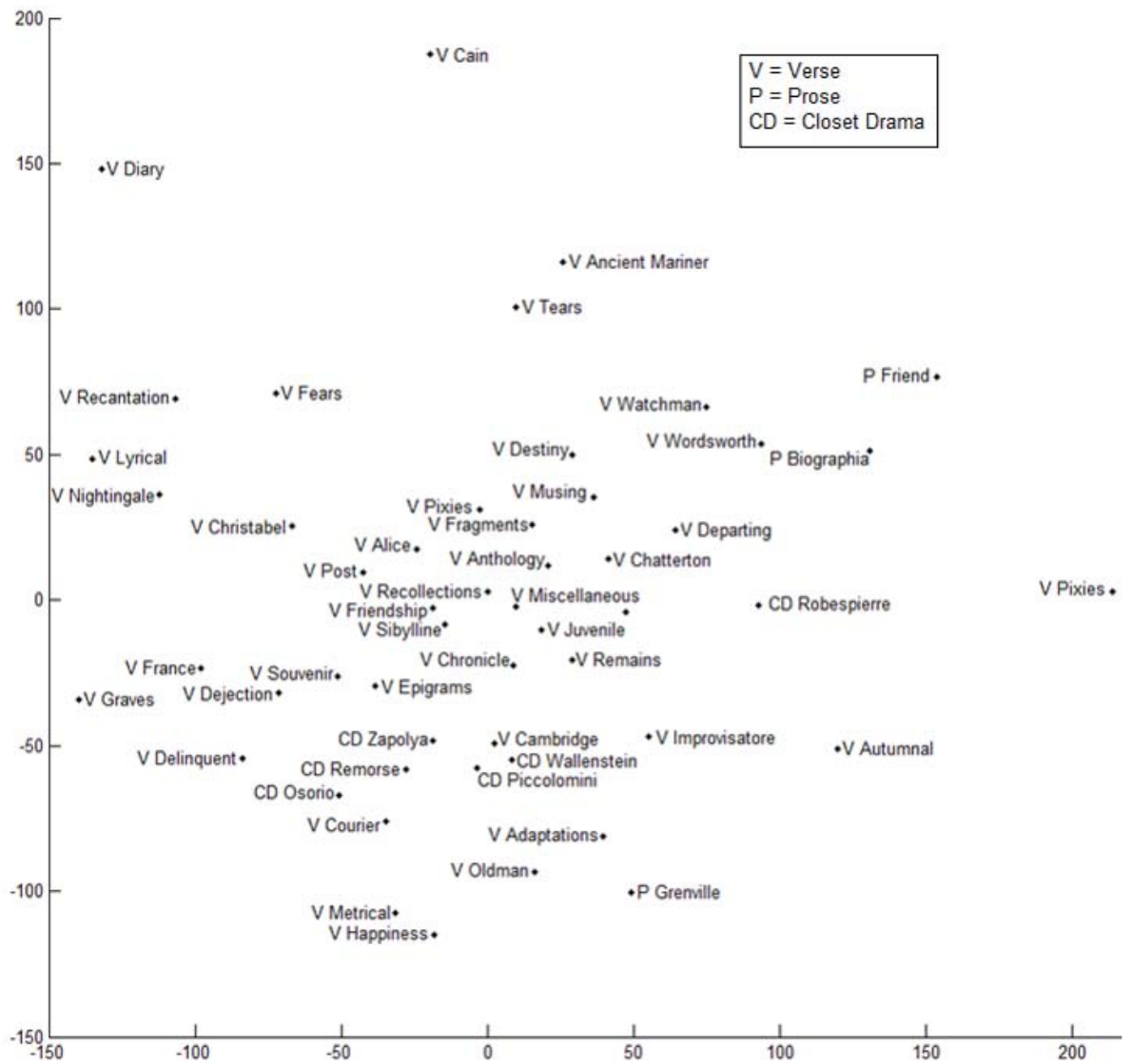


Figure (3.9) MDS of M180Norm

Isomap:

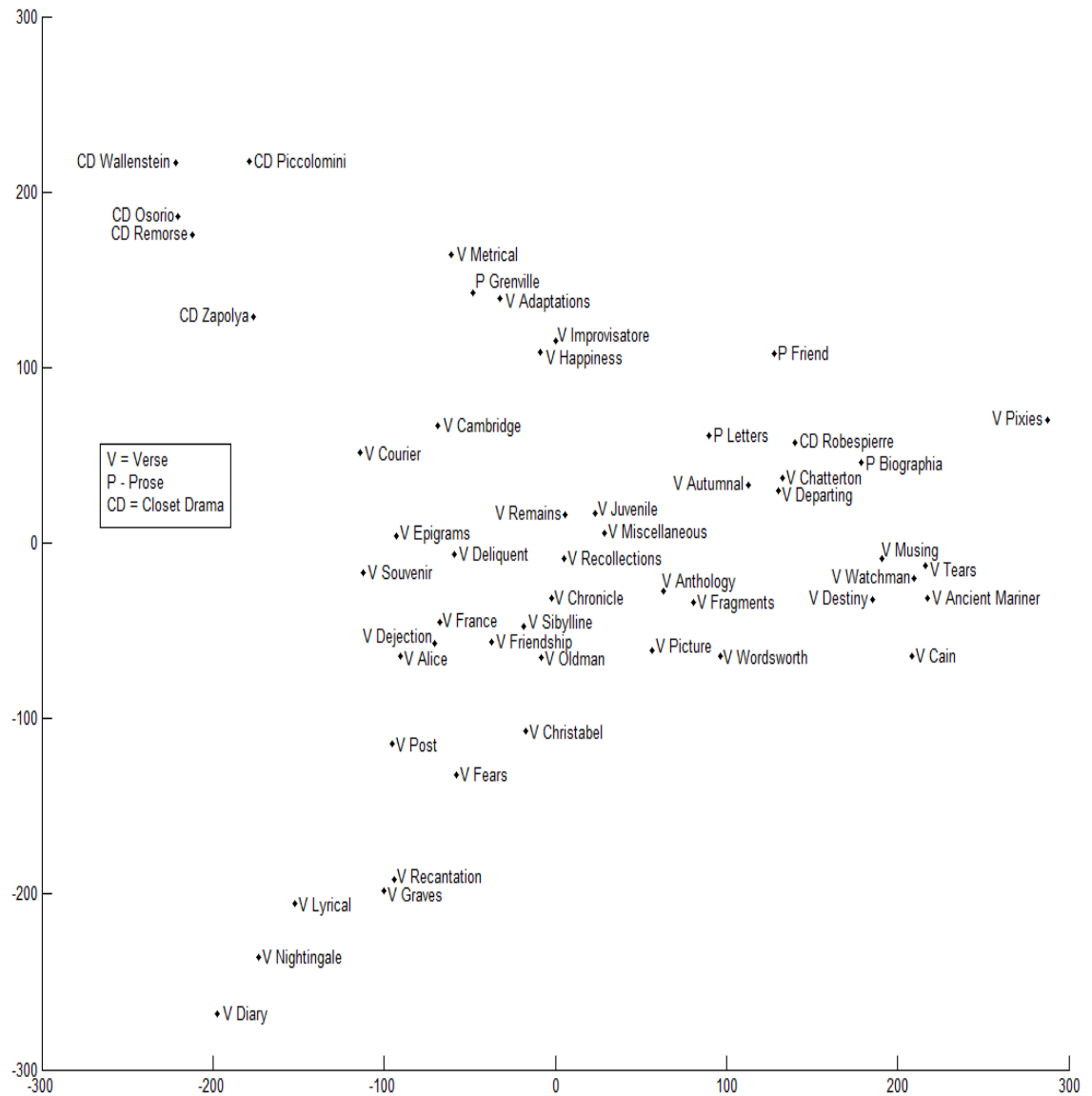


Figure (3.10) Isomap of M180Norm

SOM:

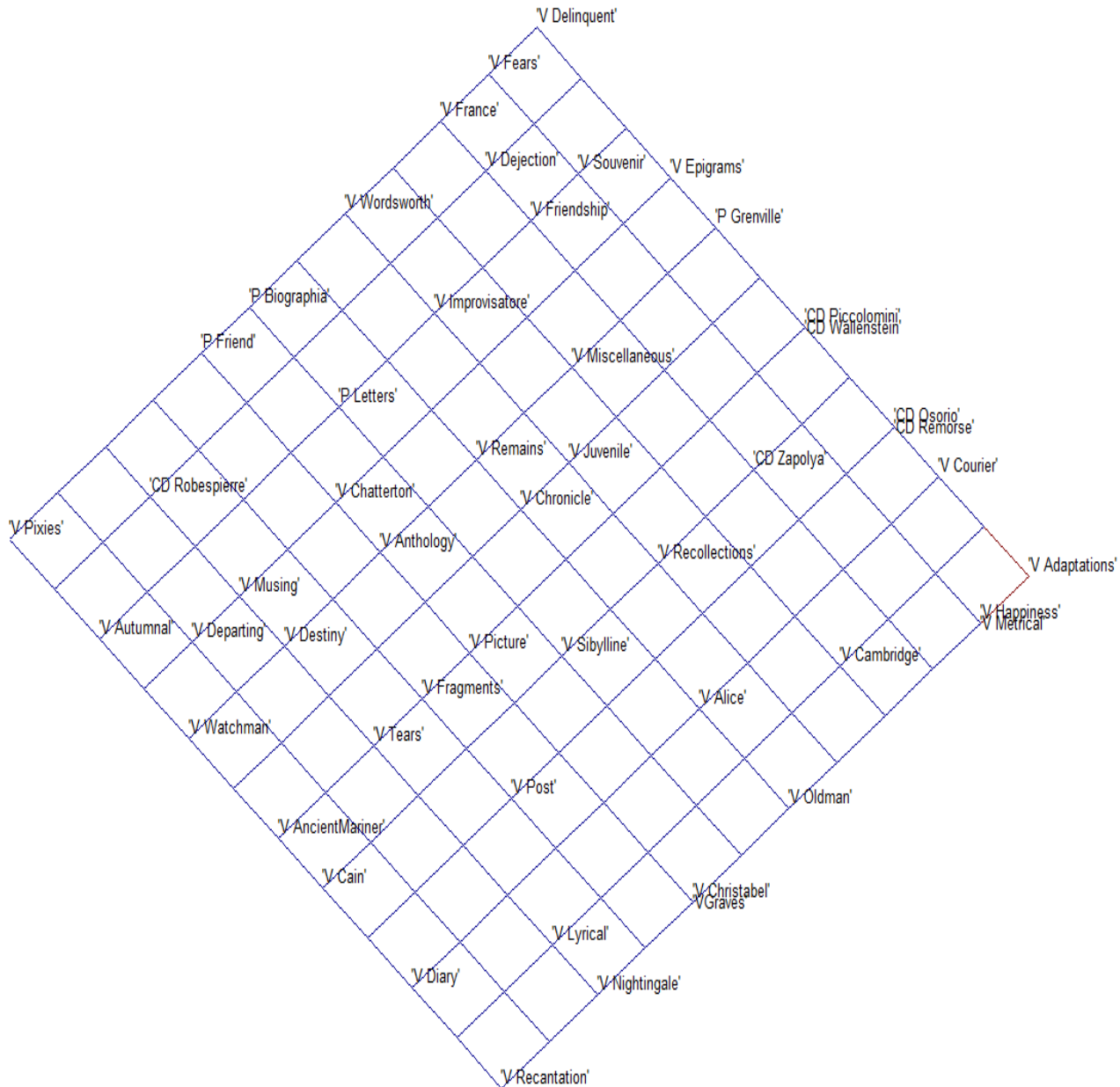


Figure (3.11) SOM of M180Norm

Despite differences of detail, the hierarchical and non-hierarchical analyses agree in clustering Coleridge's works by genre. A close observation shows that these compromise generalization about Coleridge's style in verse, prose, and drama. For example, *Zapolya*, *Remorse*, *Osorio*, *Wallenstein*, and *Piccolomini* are close to each other in one sub-cluster in the average hierarchical analysis and are also close to each other in the space generated by the non-hierarchical methods. Obviously, this is because they are all dramatic works. Similarly, *Picture*, *Letters*, *Remains*, *Juvenile*, *Recollections*, *Sibylline*, *Friendship*, *Miscellaneous*, *Fragments*, *Chronicle*, *Musing*, *Destiny*, *Chatterton*, *Anthology* are close to each other in one sub-cluster in the average hierarchical clustering and also are near

each other in the non-hierarchical methods because they are all poetical works, and so are: *Recantation, Nightingale, Lyrical, Diary; Tears, Cain, Ancient Mariner; Delinquent, Fears, France, Dejection; Metrical, Happiness, Adaptation, Improvisatore, Grenville.*

There are some inconsistencies, but these do not compromise the generalization. Examination shows that individual texts (two or more) that are placed together or close to each other in one sub-cluster by the average hierarchical method are either far from each other or any two of them are near each other in the space in one or a couple of non-hierarchical methods. Examples include the sub-cluster consisting of *Wordsworth, Friend, Biographia*; the sub-cluster consisting of *Robespierre, Watchman, Departing, and Autumnal*; the sub-cluster consisting of *Cambridge and Alice*, the sub-cluster consisting of *Post, Souvenir, and Epigrams*, and finally the sub-cluster consisting of *Oldman, Graves, and Christabel.*

The conclusion to this part of the study is therefore that there is structure in Coleridge's usage of function words: that usage varies in accordance with genre.

3.3 Comparison of Coleridge's usage of function words with that of contemporary authors:

If the fundamental assumption of authorship attribution is true, i.e., that each author has a characteristic style, then the logical expectation is that cluster analysis of Coleridge's literary output together with that of other authors will assign the various authors to separate clusters. To test this, samples of function word usage from the literary output of Coleridge's contemporaries Shelley, Byron, and Wordsworth were used as comparators. The selection from each author reflected the generic range of Coleridge's work: shorter lyrical poems, longer poems, prose, and closet dramas. Table (3.7) lists the works used.

Poet	Selected works
Byron	<i>Cain: A mystery, The Deformed Transformed, The two Foscari, Child Harold's Pilgrimage, Heaven and Earth, a selection of letters, Manfred: a dramatic poem, a selection of shorter poems, Werner; or, The Inheritance, Sardanapalus</i>
Shelley	<i>Adonias: An elegy on the death of John Keats, The Cenci, A defence of poetry and other essays, Faust, Prometheus Unbound, A selection of</i>

	shorter poems
Wordsworth	<i>The Borderers</i> , a selection of letters, <i>The prelude</i> , a selection of shorter poems, poetry as a study

Table (3.7) A selection of works from Byron, Shelley, and Wordsworth

As before, we used an abbreviation for each of these works to refer to either work by any one of Coleridge’s contemporaries across all five analyses.

These texts were pre-processed as for those of Coleridge, described earlier, to remove extraneous additions and then added to the above works by Coleridge to constitute a new corpus. A function word frequency matrix F2 was abstracted from this enlarged corpus, length-normalized as above and dimensionality-reduced to the 80 most frequent words in accordance with figure (3.12).

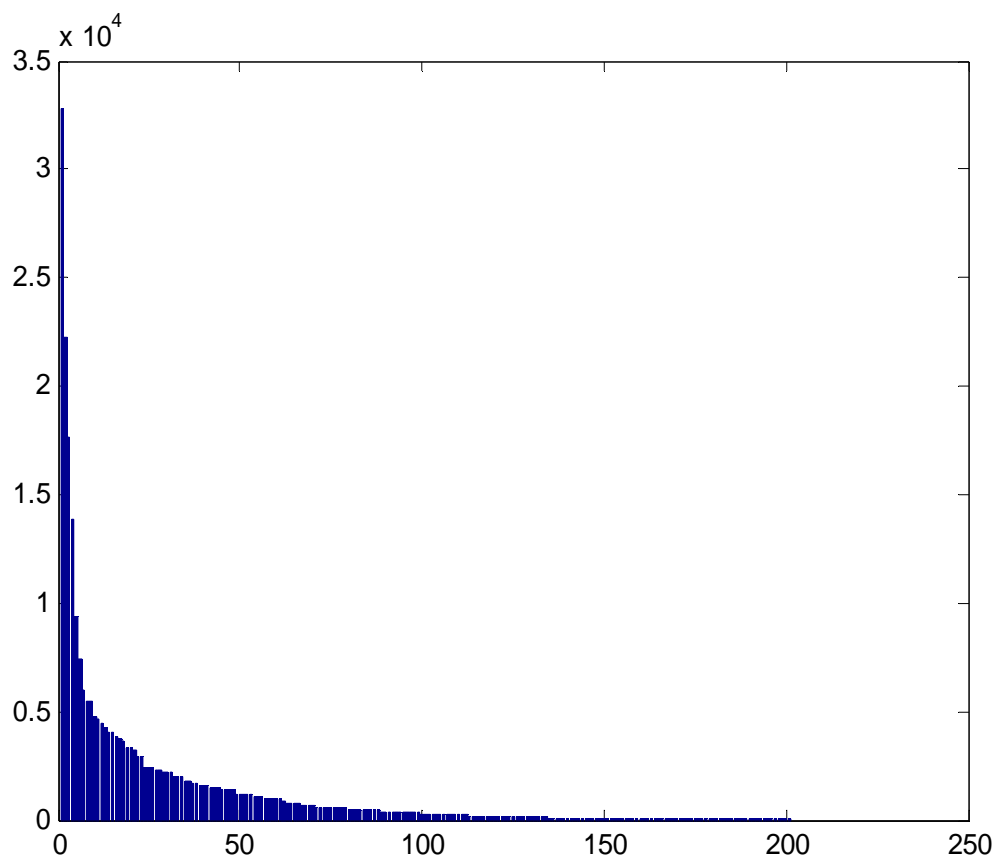


Figure (3.12) The distribution of function words in frequency matrix F2

The resulting matrix M280Norm was then cluster analysed using the same methods as those applied to the Coleridge-only corpus, with the following results.

Single Linkage (Cophenetic correlation: 0.7201):

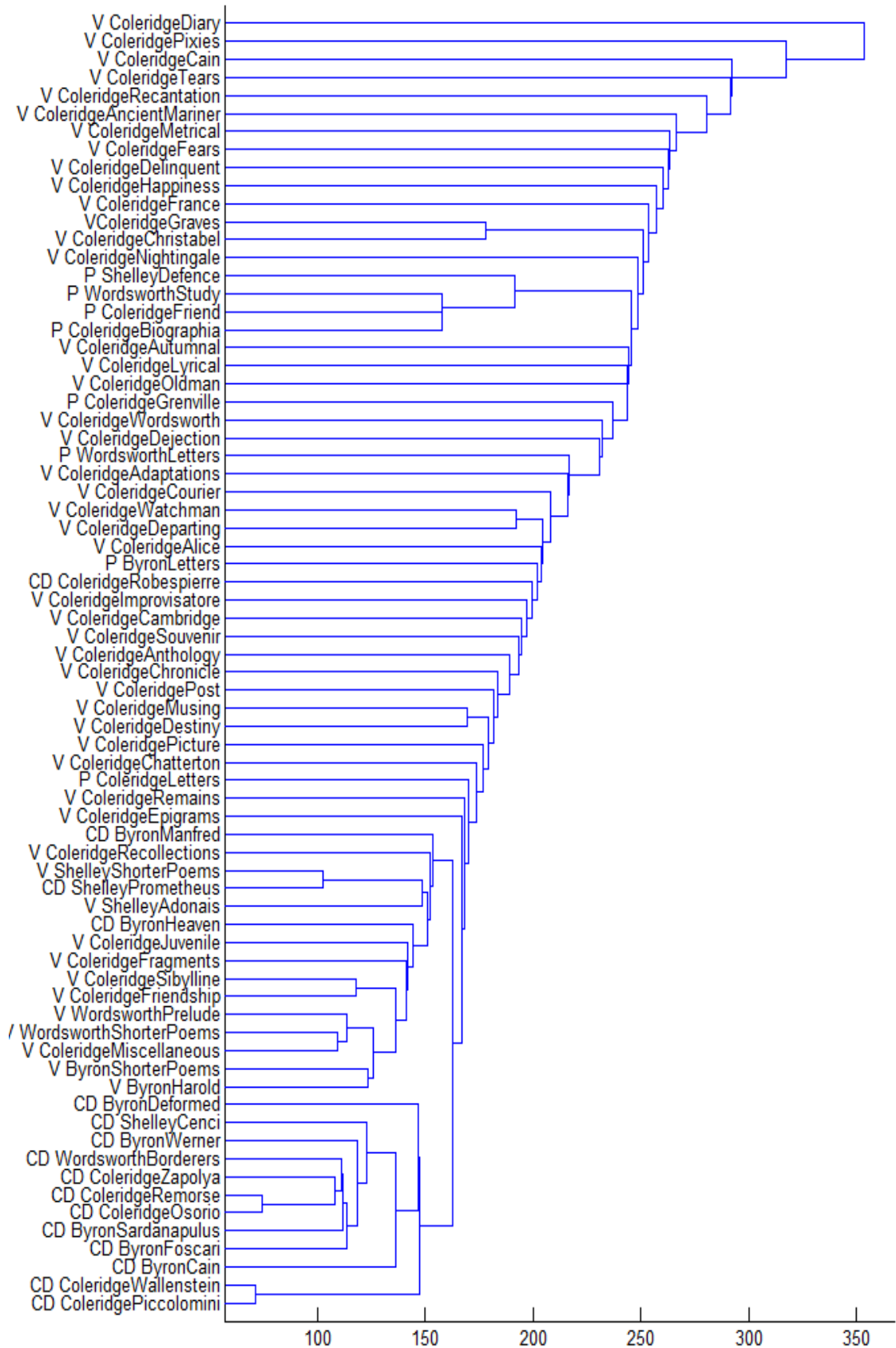


Figure (3.13): Single Linkage. Cophenetic correlation: 0.7201

Complete Linkage (Cophenetic correlation: 0.6947):

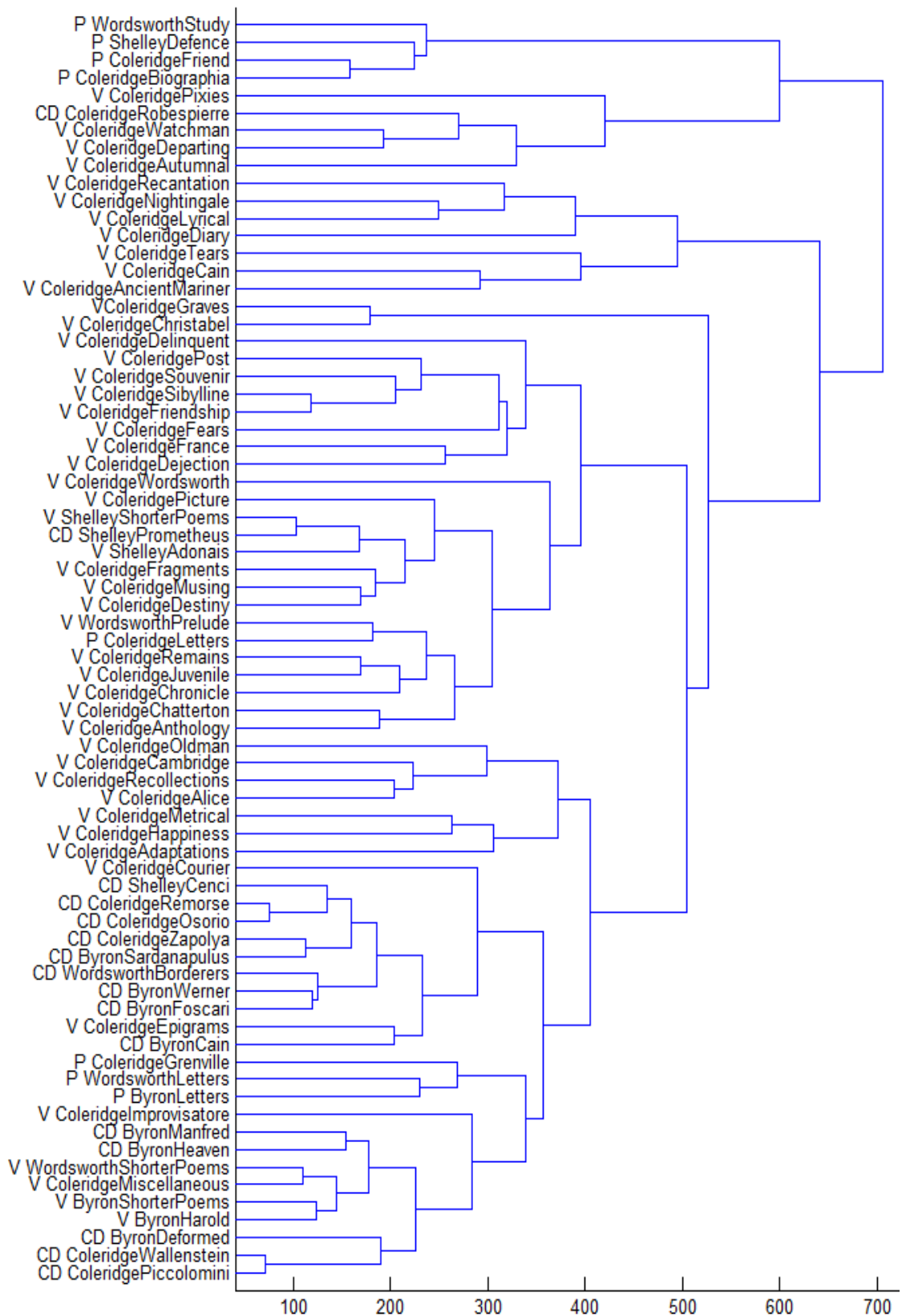


Figure (3.14): Complete Linkage. Cophenetic correlation: 0.6947

Average Linkage (Cophenetic correlation: 0.7705):

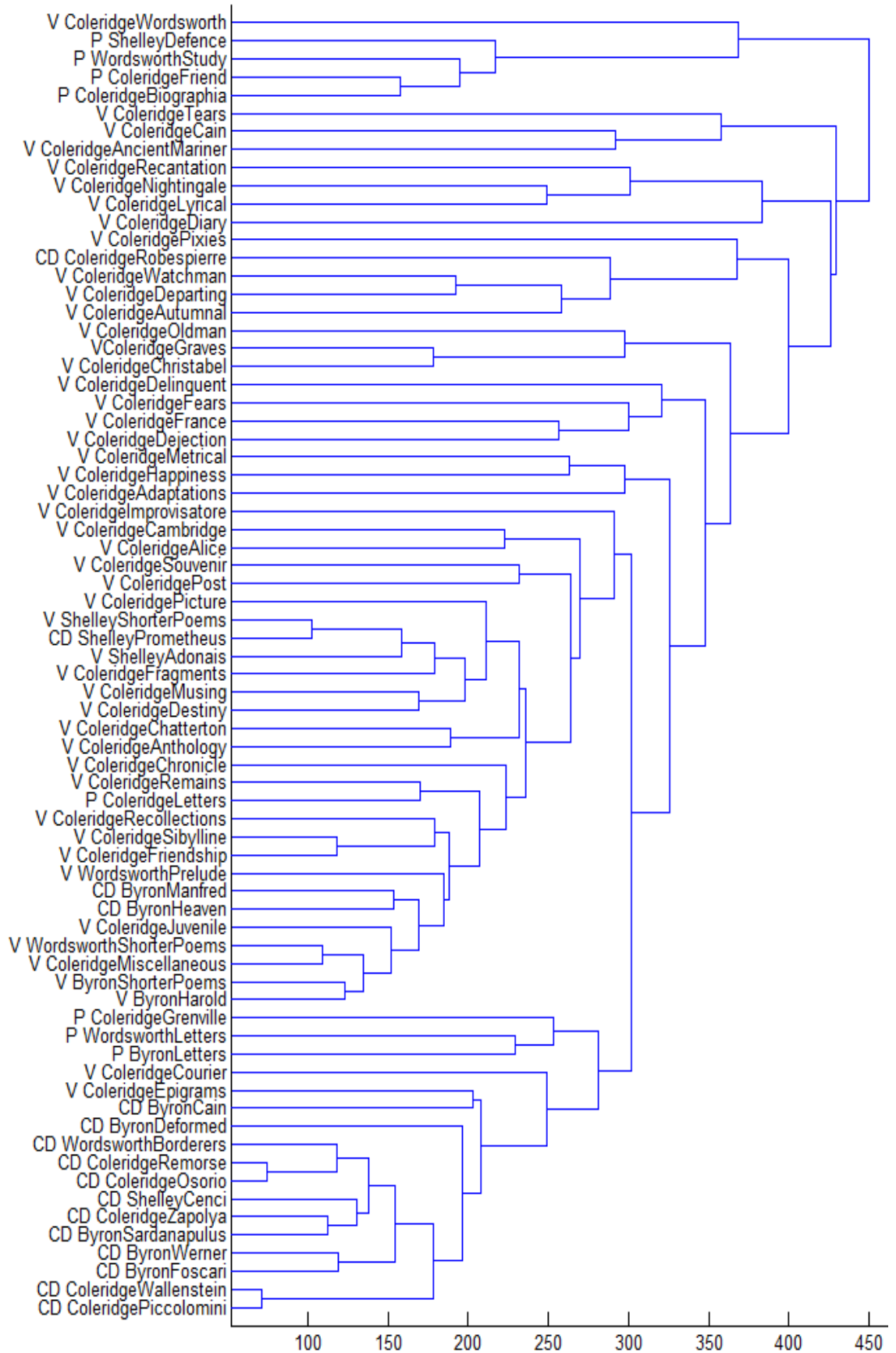


Figure (3.15): Average Linkage. Cophenetic correlation: 0.7705

Ward linkage (Cophenetic correlation: 0.4356):

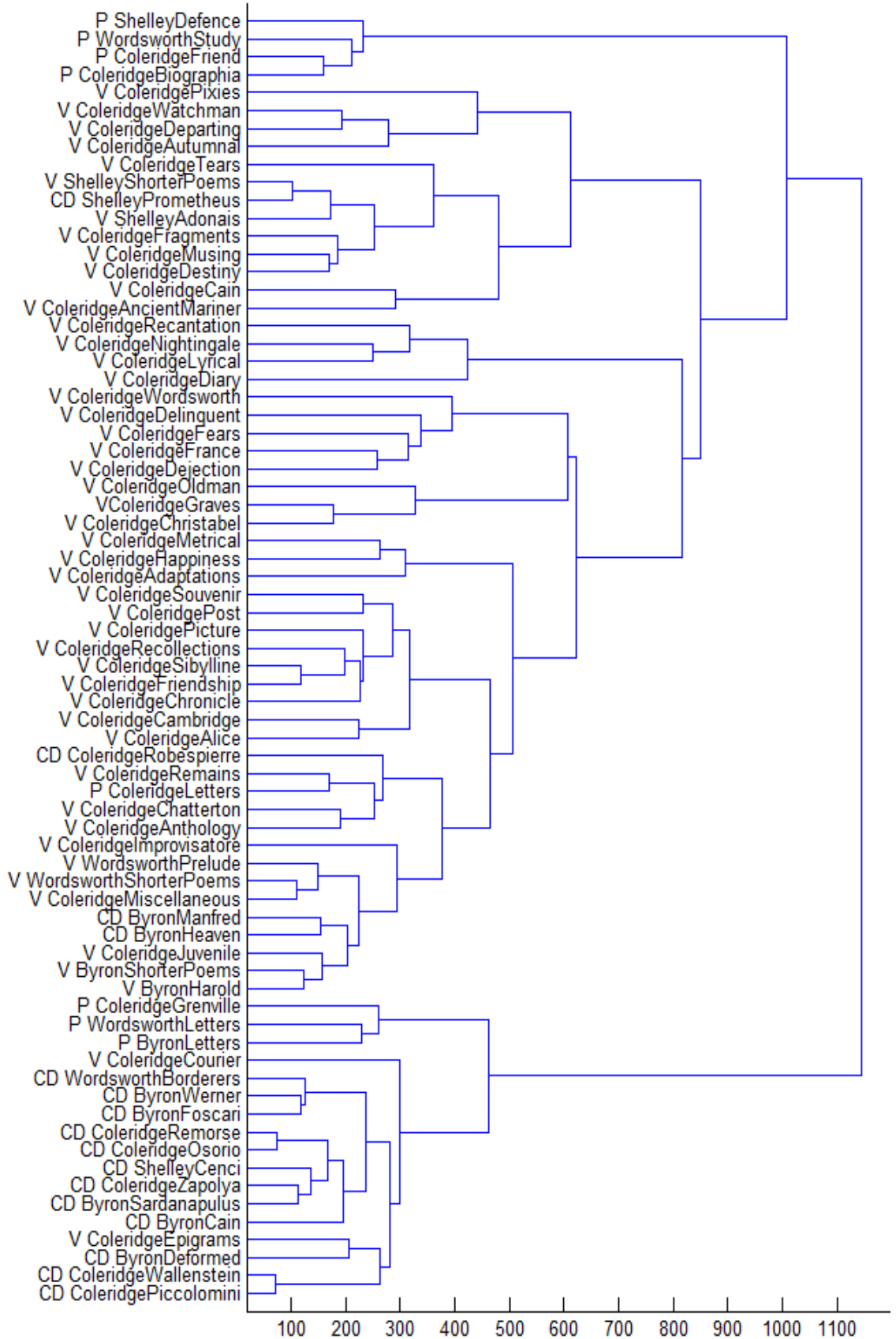


Figure (3.16): Ward linkage. Cophenetic correlation: 0.4356

Average linkage clustering for M280Norm is again best for the cophenetic correlation coefficient criterion, as shown in Table (3.8).

Hierarchical clustering method	Cophenetic correlation coefficient
Single	0.7201
Complete	0.6947
Average	0.7705
Ward	0.4356

Table (3.8) Cophenetic correlation coefficients for Figures (3-13, 14, 15, 16)

As before, these hierarchical results were validated by comparison with results from non-hierarchical clustering methods.

A general problem with non-hierarchical methods is that, as the number of objects being clustered increases, the labelling tends to obscure the underlying structure. The labelled non-hierarchical results are therefore preceded by unlabelled ones to show the underlying structure for PCA, MDS, and Isomap; SOM remains clear with labelling and is not included.

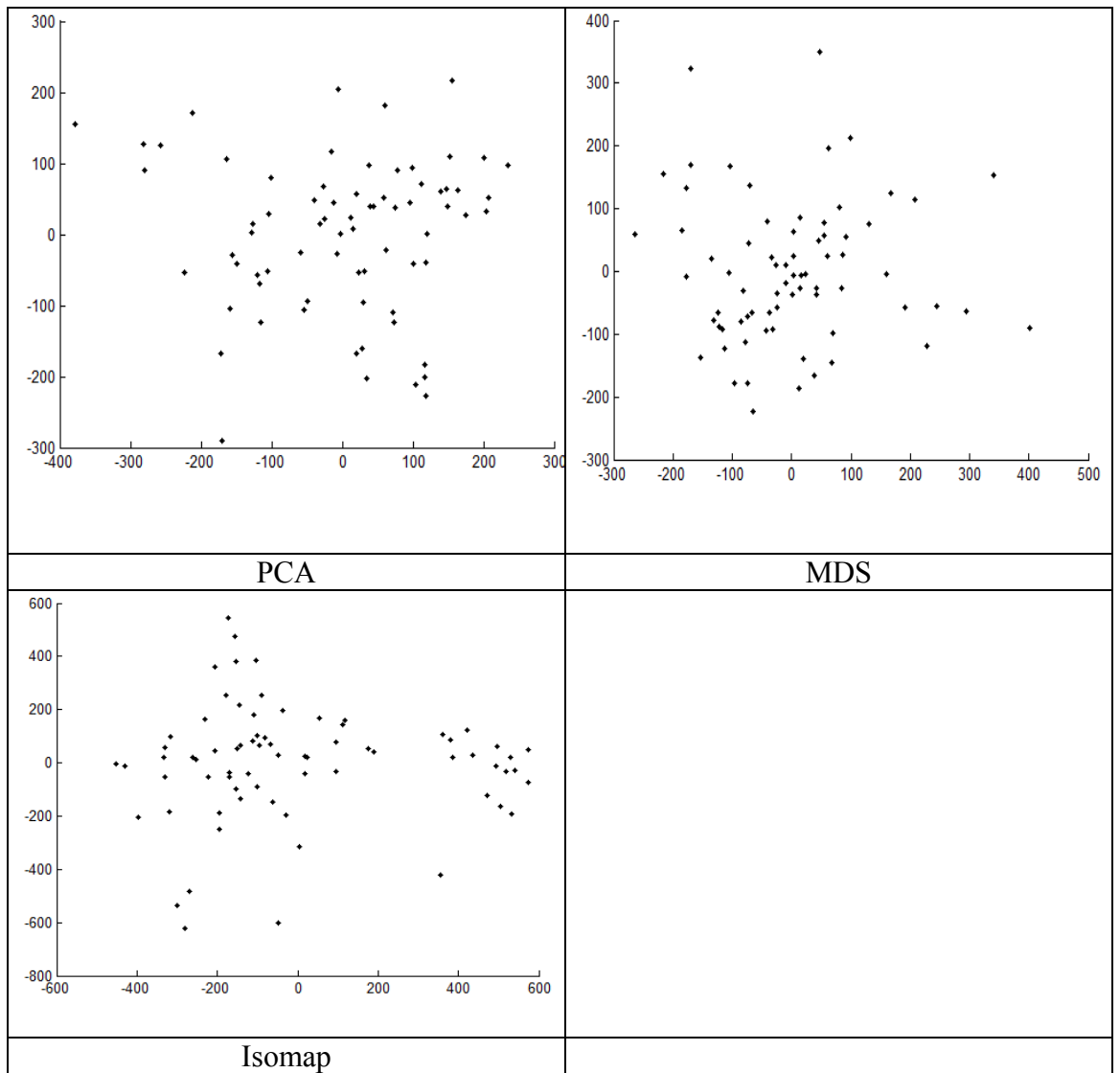


Figure (3.17): Unlabelled clustering results

PCA:

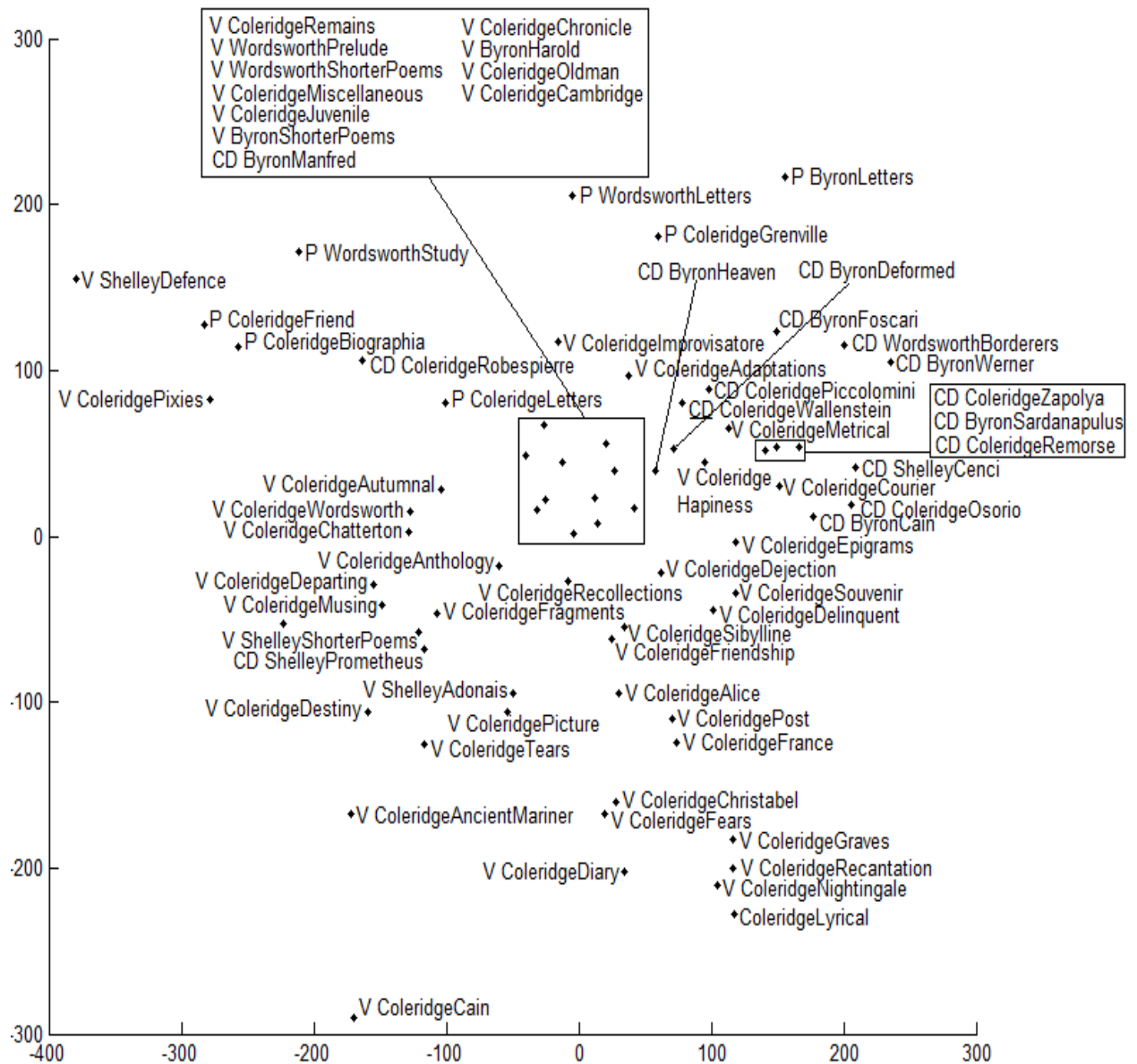


Figure (3.18): PCA of M280Norm

MDS:

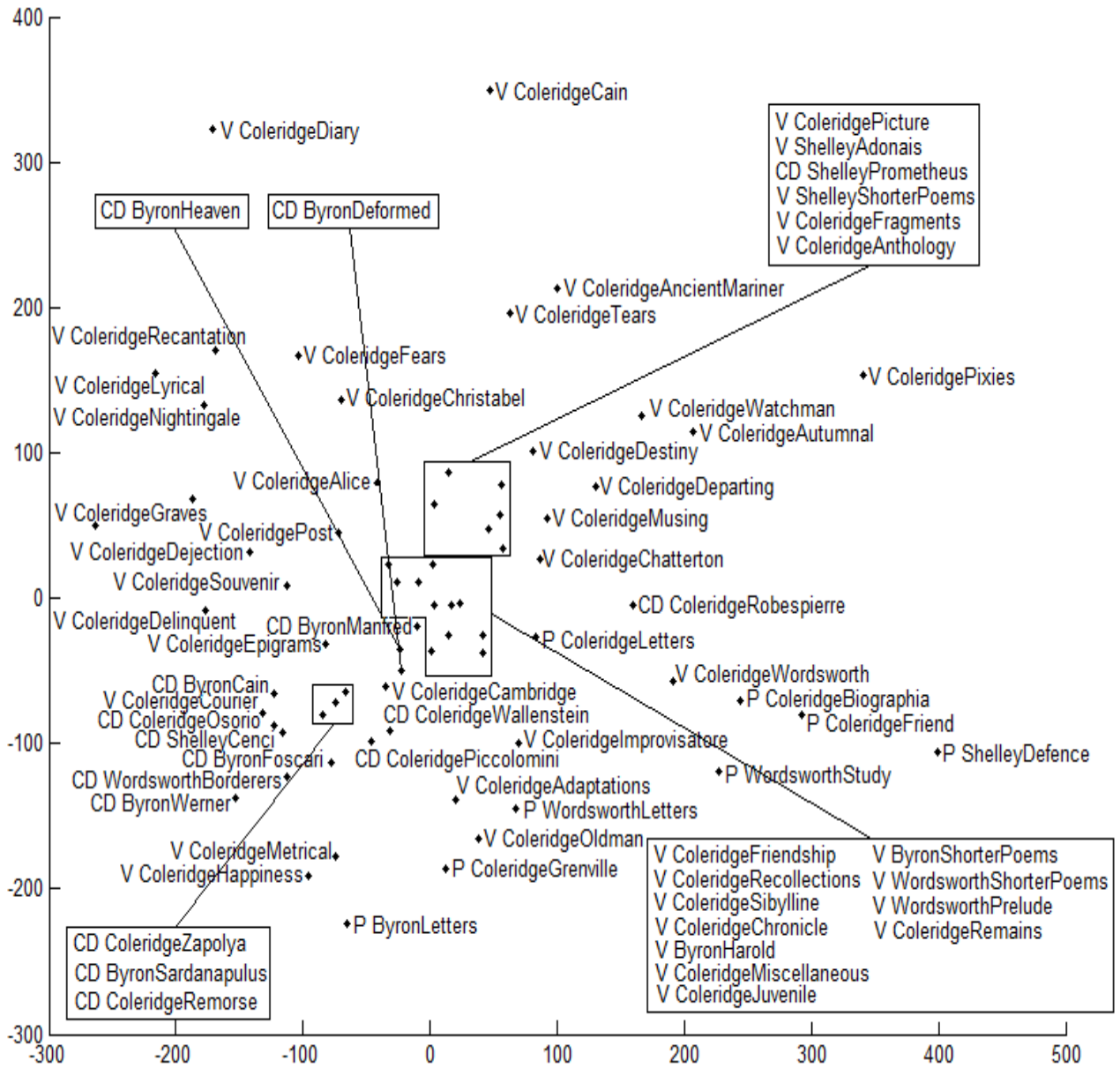


Figure (3.19): MDS of M280Norm

Isompa:

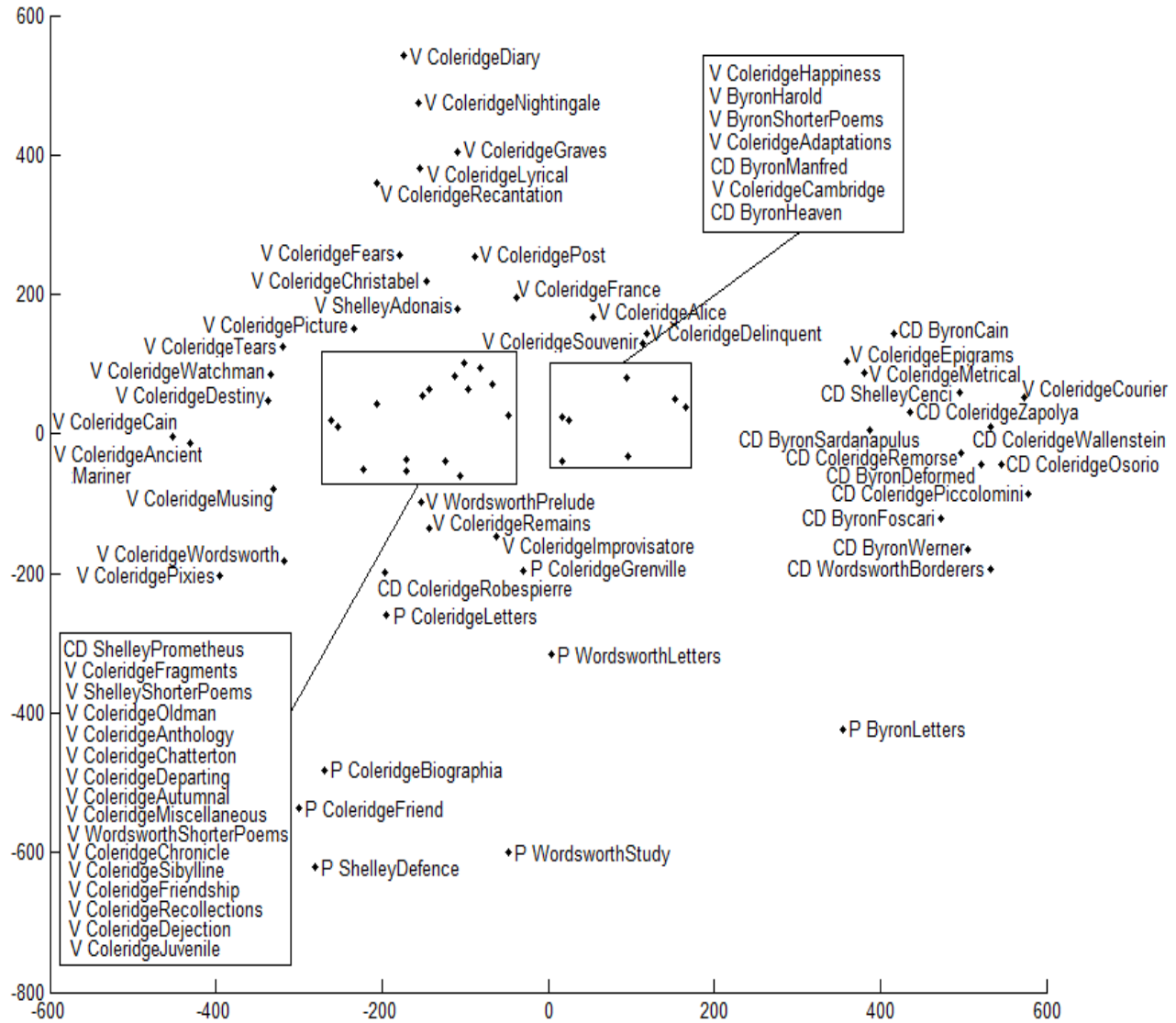


Figure (3.20): Isomap of M280Norm

SOM:

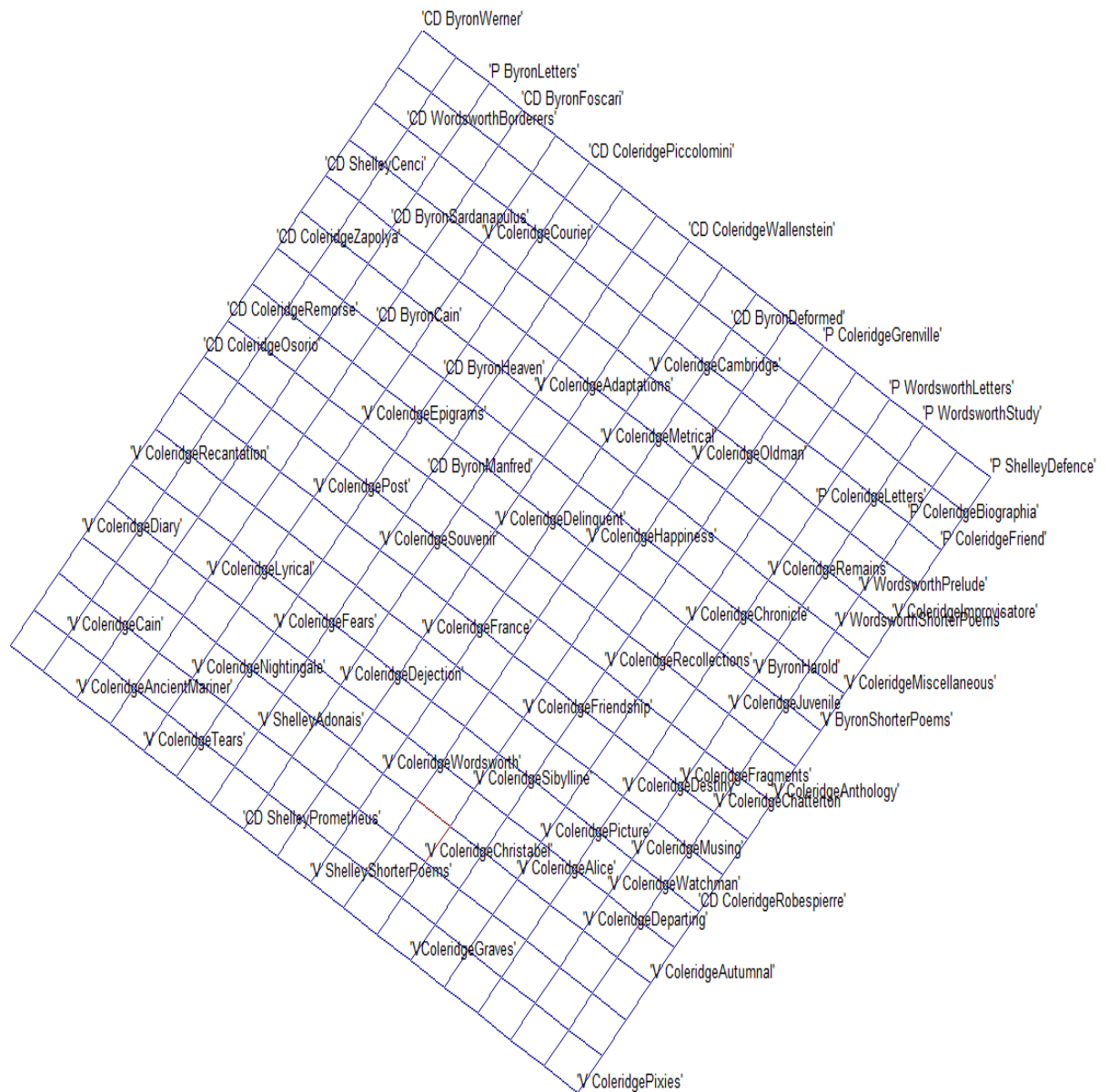


Figure (3.21) SOM of M280Norm

Comparison of the results of the five clustering methods applied to the corpus of 72 texts supports the following conclusions:

- As with the analyses of section (3.2) above, the hierarchical and non-hierarchical results agree. The texts that are close to other texts in any sub-clusters generated by the average hierarchical are also close, with few exceptions, to each other in the non-hierarchical methods. Examples include, the sub-cluster consisting of Coleridge *Epigrams*, Byron *Cain*, Byron *Deformed*, Wordsworth *Borderers*, Coleridge *Remorse*, Coleridge *Osorio*, Shelley *Cenci*, Coleridge *Zapolya*, Byron *Sardanapalus*, Byron

Werner, Byron *Foscari*, Coleridge *Wallenstein*, and Coleridge *Piccolomini*; the sub-cluster consisting of Coleridge *Picture*, Shelley Shorter poems, Shelley *Prometheus*, Shelley *Adonais*, Coleridge *Fragments*, Coleridge *Musing* Coleridge *Destiny*, Coleridge *Chatterton*, Coleridge *Anthology*, Coleridge *Chronicle*, Coleridge *Remains*, Coleridge *Letters*, Coleridge *Recollections*, Coleridge *Sibylline*, Coleridge *Friendship*, Wordsworth *Prelude*, Byron *Manfred*, Byron *Heaven*, Coleridge *Juvenile*, Wordsworth Shorter Poems, Coleridge *Miscellaneous*, Byron Shorter Poems, and Byron *Harold*; the sub-cluster consisting of Coleridge *Wordsworth*, Shelley *Defence*, Wordsworth *Study*, Coleridge *Friend*, and Coleridge *Biographia*; the sub-cluster consisting of Coleridge *Grenville*, Wordsworth Letters, and Byron Letters; and the sub-cluster consisting of Coleridge *Oldman*, Coleridge *Graves*, Coleridge *Christabel*; and the sub-cluster consisting of Coleridge *Delinquent*, Coleridge *Fears*, Coleridge *France*, and Coleridge *Dejection*; the sub-cluster consisting of Coleridge *Cambridge* and Coleridge *Alice*, and finally the sub-cluster consisting of Coleridge *Post*, and Coleridge *Souvenir*.

- Also as with the analyses of section (3.2), clustering is by literary genre, with verse, prose, and closet drama forming their own distinct clusters. However, the clustering results show that some individual texts that are close to each other in one sub-cluster in the average hierarchical method are either far away from each other in the space or are located near each other but one or two texts are far apart in one or a couple of non-hierarchical methods. Examples include: Coleridge *Metrical*, Coleridge *Happiness*, Coleridge *Adaptations* and Coleridge *Improvisatore*; Coleridge *Tears*, Coleridge *Cain*, Coleridge *Ancient Mariner*, Coleridge *Recantation*, Coleridge *Nightingale*, Coleridge *Lyrical*, and Coleridge *Diary*; Coleridge *Pixies*, Coleridge *Robespierre*, Coleridge *Watchman*, Coleridge *Departing*, and Coleridge *Autumnal*; Coleridge *Oldman*; Wordsworth *Prelude*; and finally Coleridge *Cain*.
- Within the three generic clusters generated by the average hierarchical clustering there is no clear sub-clustering according to author, apart from the sub-cluster consisting of Coleridge's texts in verse: Coleridge *Graves*, Coleridge *Christabel*, Coleridge *Delinquent*, Coleridge *Fears*, Coleridge *France*, Coleridge *Dejection*, Coleridge *Metrical*, Coleridge *Happiness*, Coleridge *Adaptations*, Coleridge *Improvisatore*, Coleridge *Cambridge*, Coleridge *Alice*, Coleridge *Souvenir*, and Coleridge *Post*.

This result has serious implications for the validity of the central tenet of authorship

attribution; clearly, clustering of the work of a much larger range of authors is required to draw any firm conclusions about this, but the results just presented are not encouraging. More is said about this in subsequent discussion.

3.4 Where *Faustus* fits:

The next step is to see where cluster analysis places the 1821 Boosey *Faustus* in the corpus of texts by Coleridge, Shelley, Byron, and Wordsworth. The Boosey *Faustus* was pre-processed and inserted into the existing corpus, a new function word frequency matrix F3 was extracted, and F3 was length-normalized and dimensionality-reduced to 80 as before. The clustering results follow.

Single Linkage (Cophenetic correlation coefficient: 0.7235):

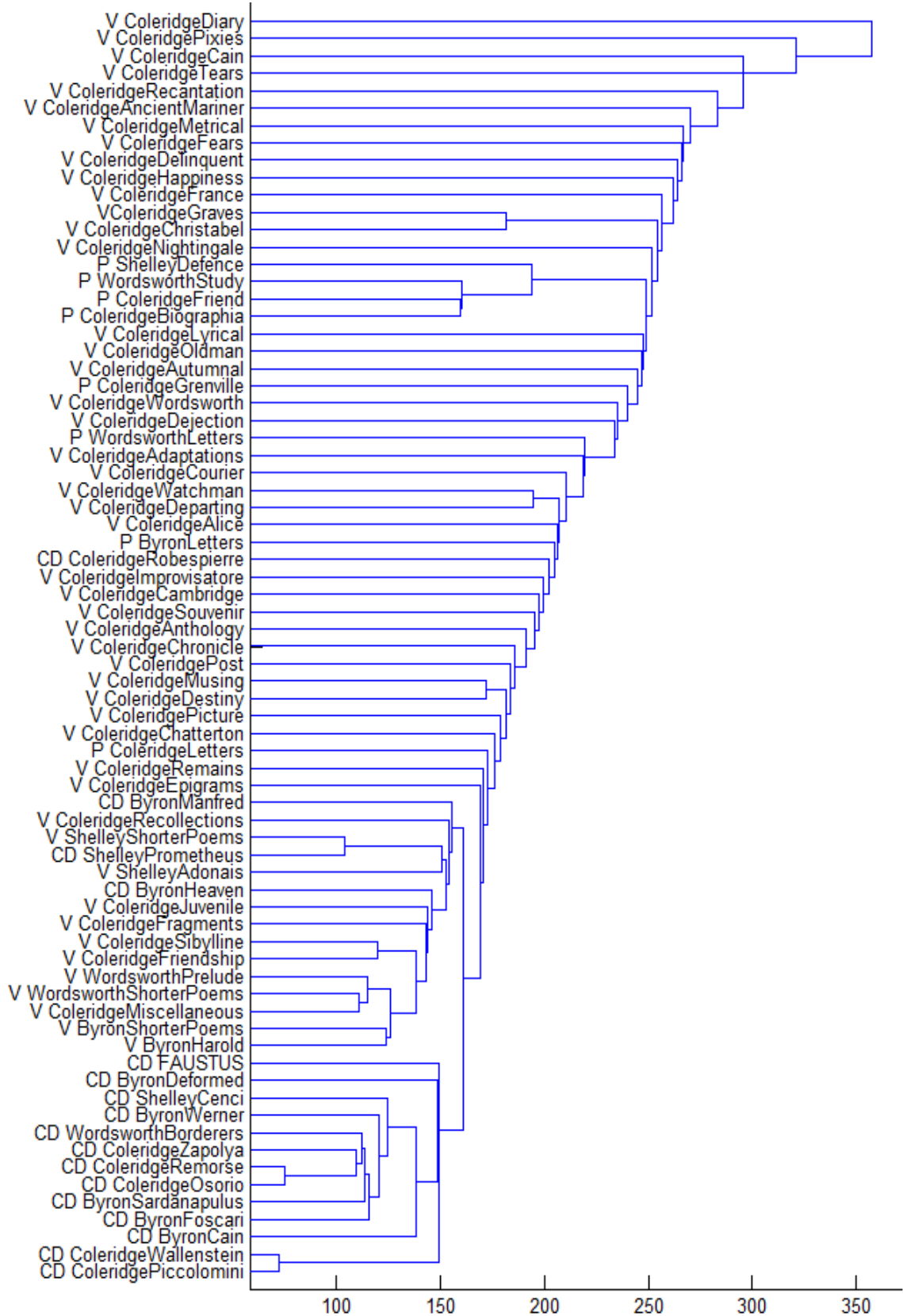


Figure (3.22) Single Linkage. Cophenetic correlation coefficient: 0.7235

Complete Linkage (Cophenetic correlation coefficient: 0.6978):

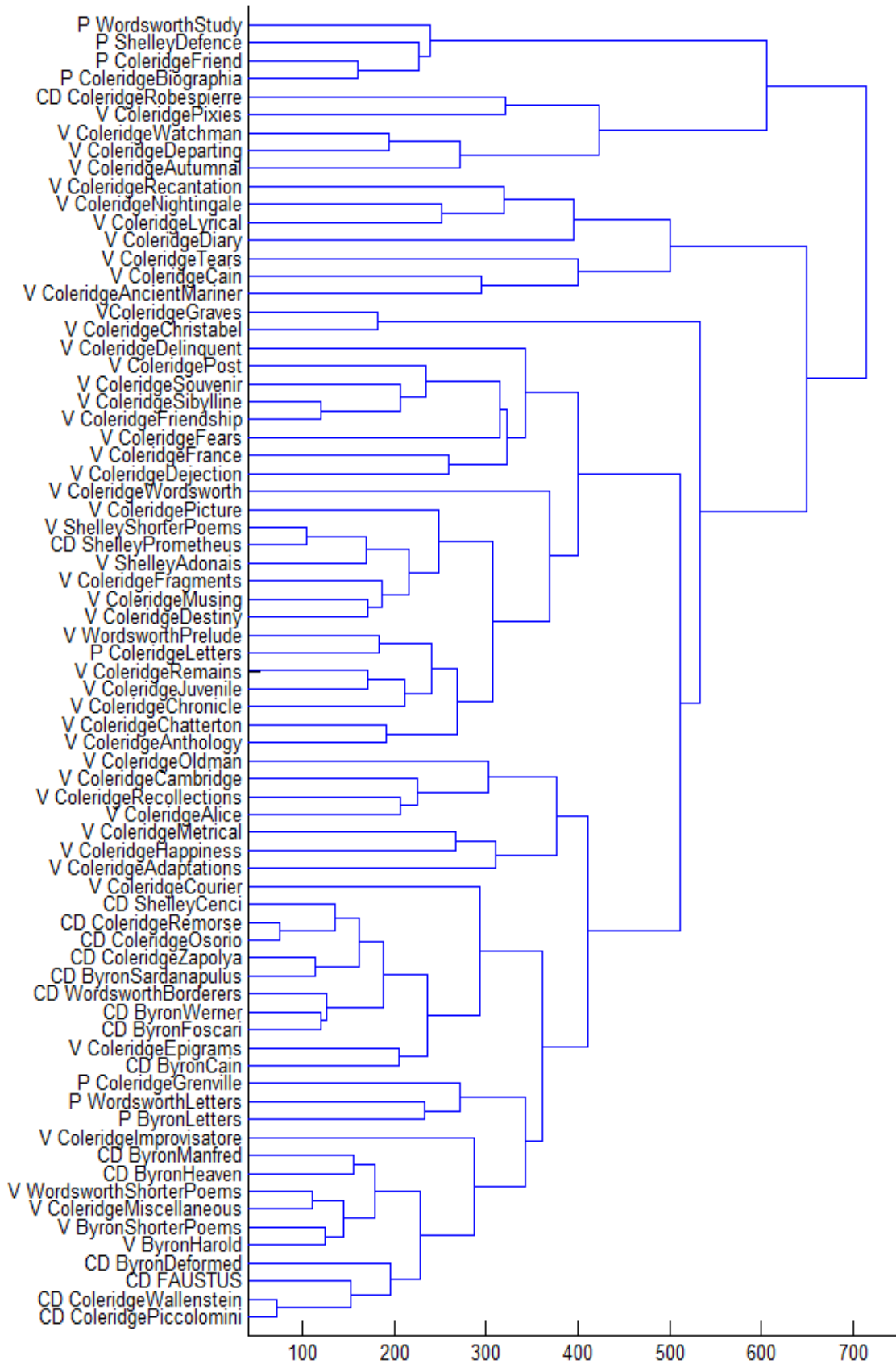


Figure (2.23) Complete Linkage. Cophenetic correlation coefficient: 0.6978

Average Linkage (Cophenetic correlation coefficient: 0.7732):

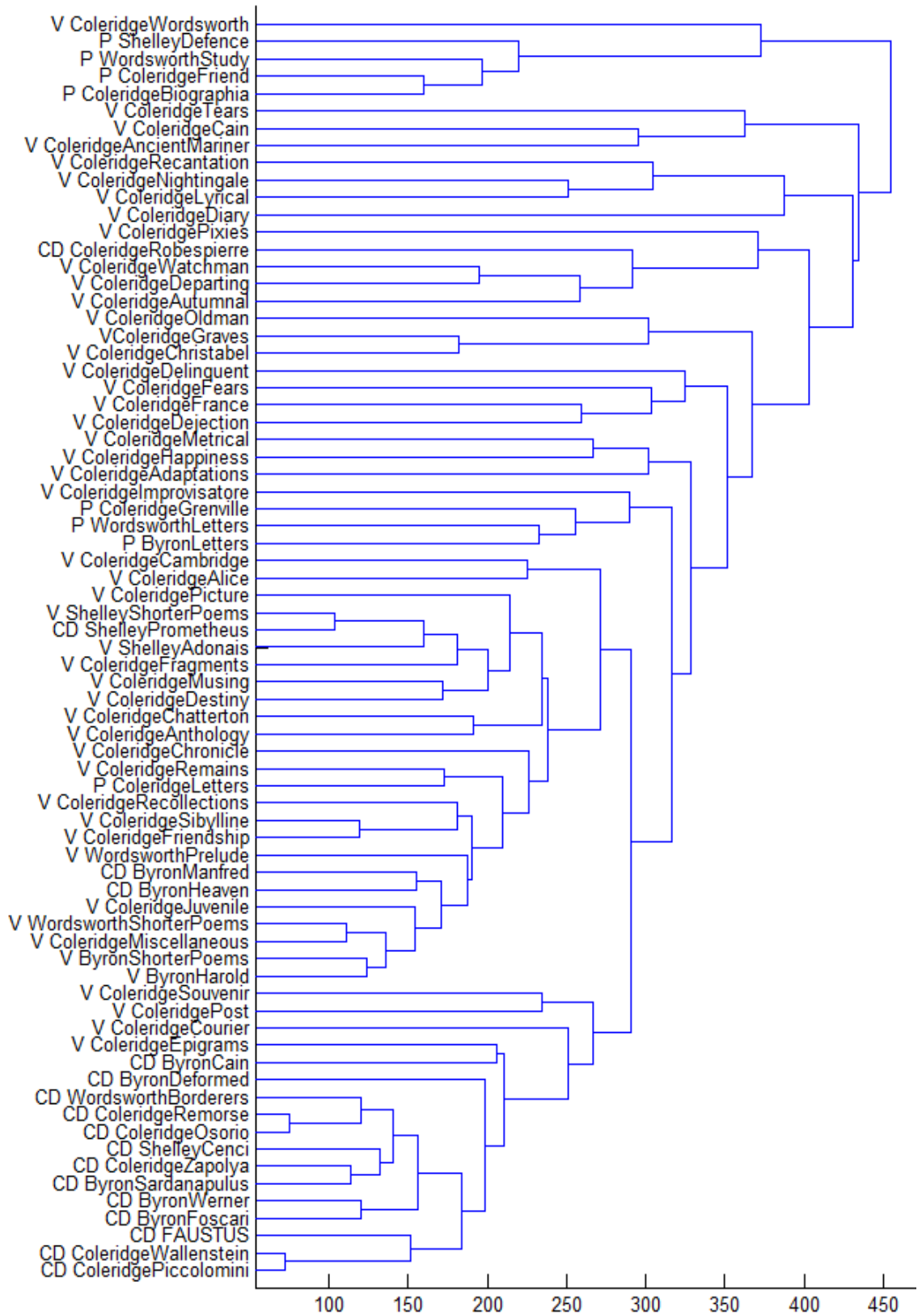


Figure (3.24) Average Linkage. Cophenetic correlation coefficient: 0.7732

Ward Linkage (Cophenetic correlation coefficient: 0.4244):

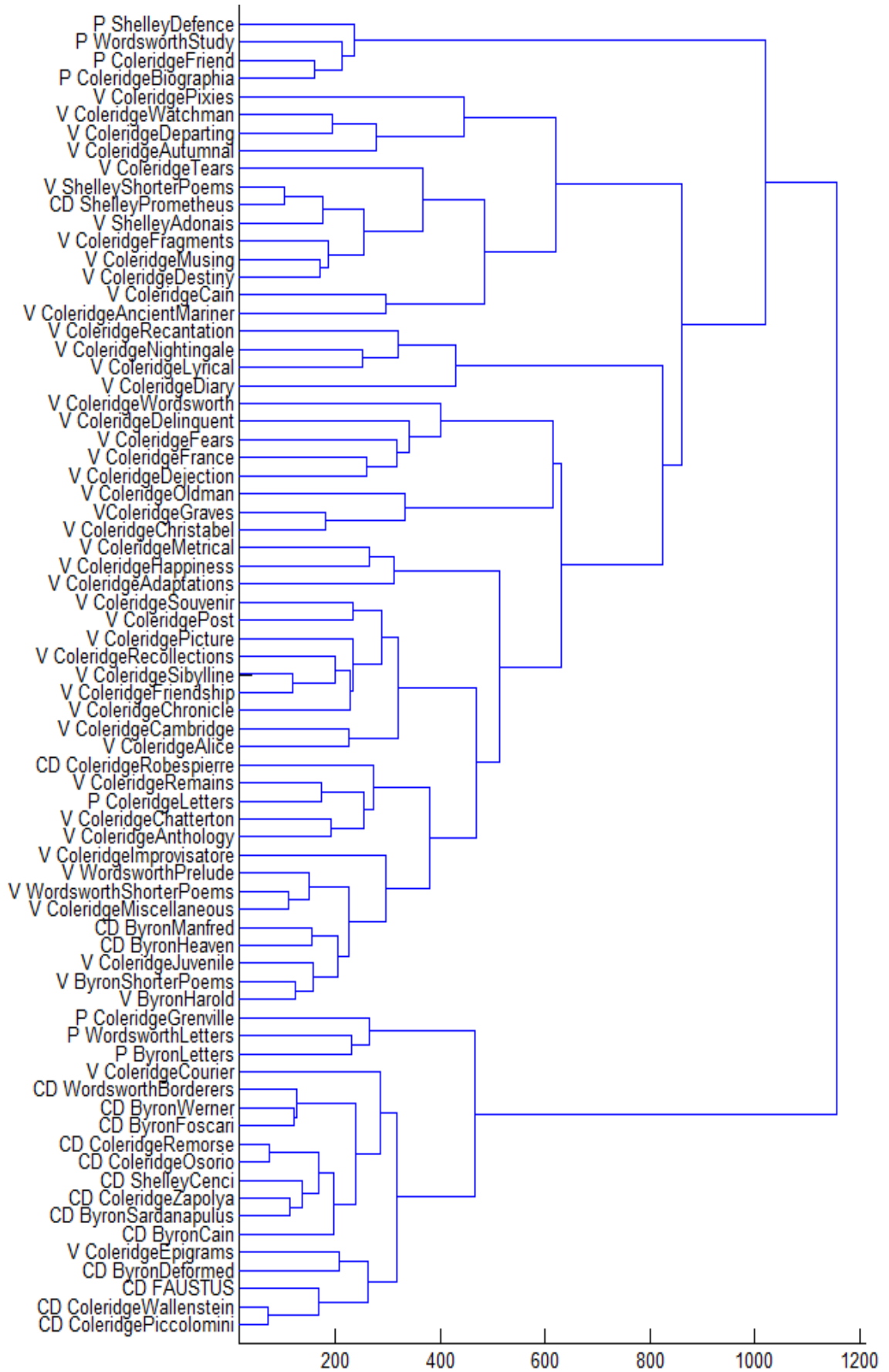


Figure (3.25) Ward Linkage. Cophenetic correlation coefficient: 0.4244

Based on the validation by cophenetic correlation coefficient, the hierarchical clustering tree generated by Average clustering analysis seems to fit M380Norm data matrix more better than the clusterings produced by Single, Complete, and Ward analyses.

Hierarchical clustering method	Cophenetic correlation coefficient
Single	0.7235
Complete	0.6978
Average	0.7732
Ward	0.4244

Table (3.9) Cophenetic correlation coefficient for matrix M380Norm

Again, another validation is by non-hierarchical clustering methods. Since the overall cluster structure is known from the immediately preceding section, only the texts in the immediate neighbourhood of *Faustus* are labelled to avoid overloading and thereby obscuring the cluster results.

PCA:

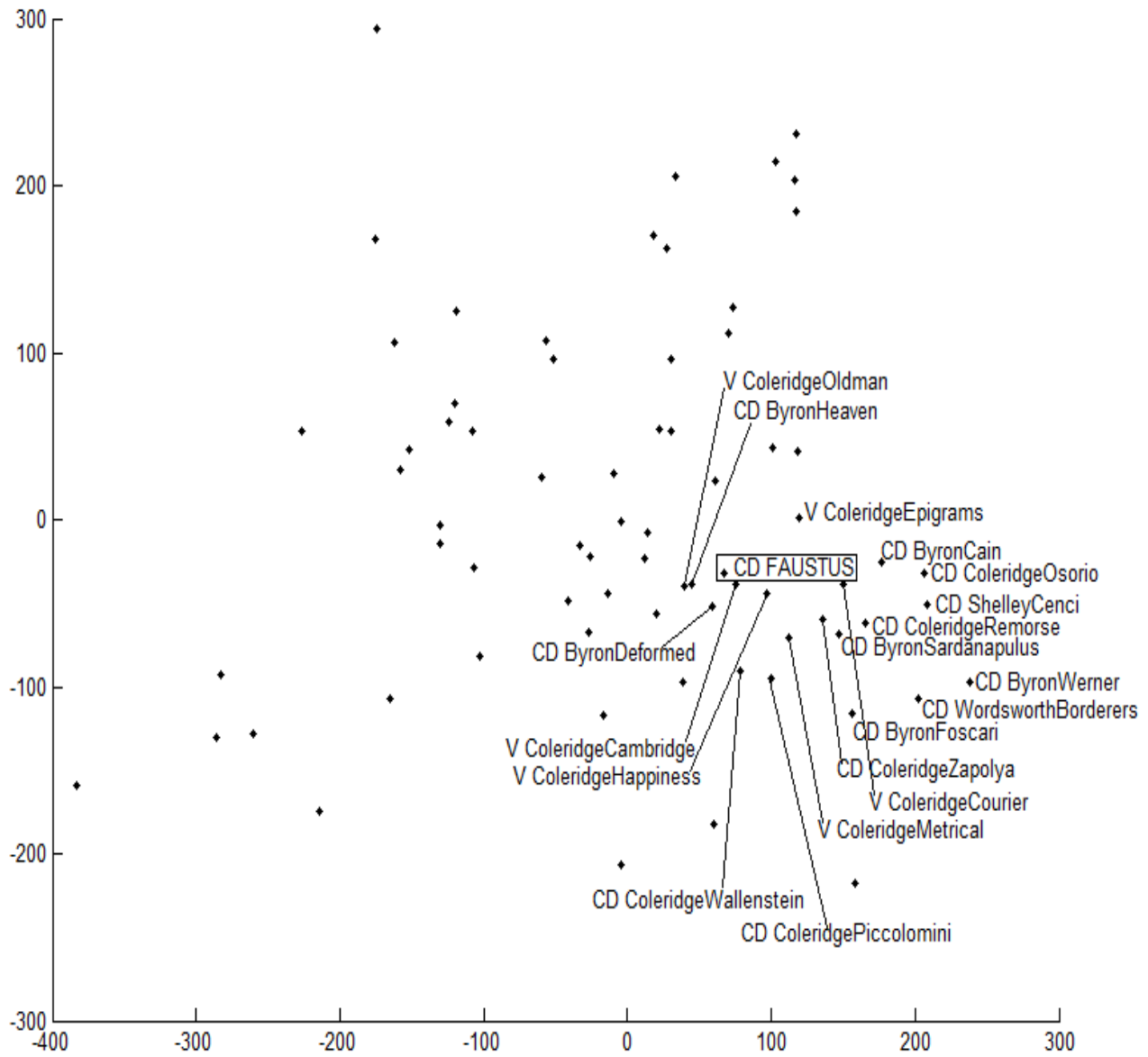


Figure (3.26) PCA of M380Norm

MDS:

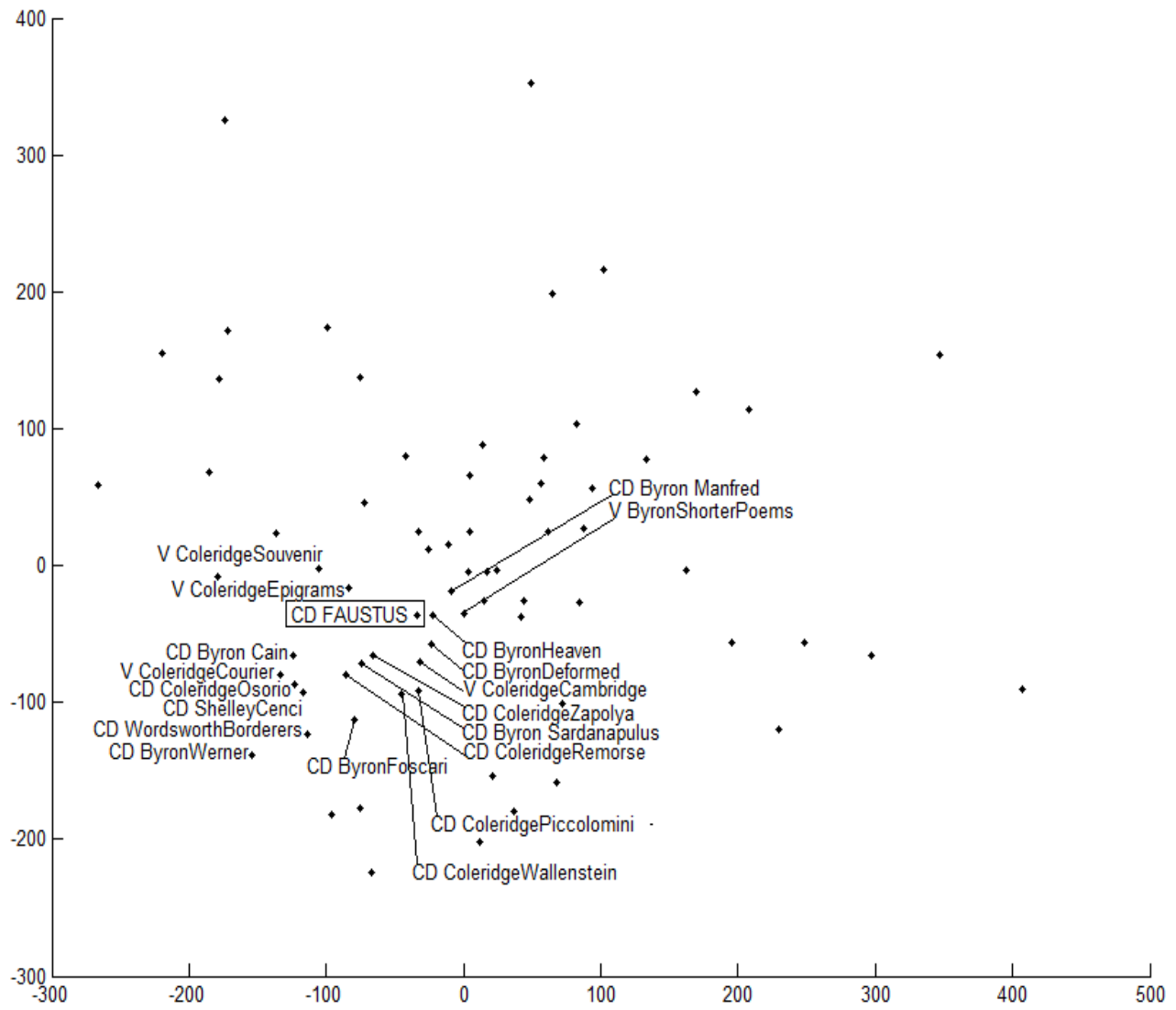


Figure (3.27) MDS of M380Norm

Isomap:

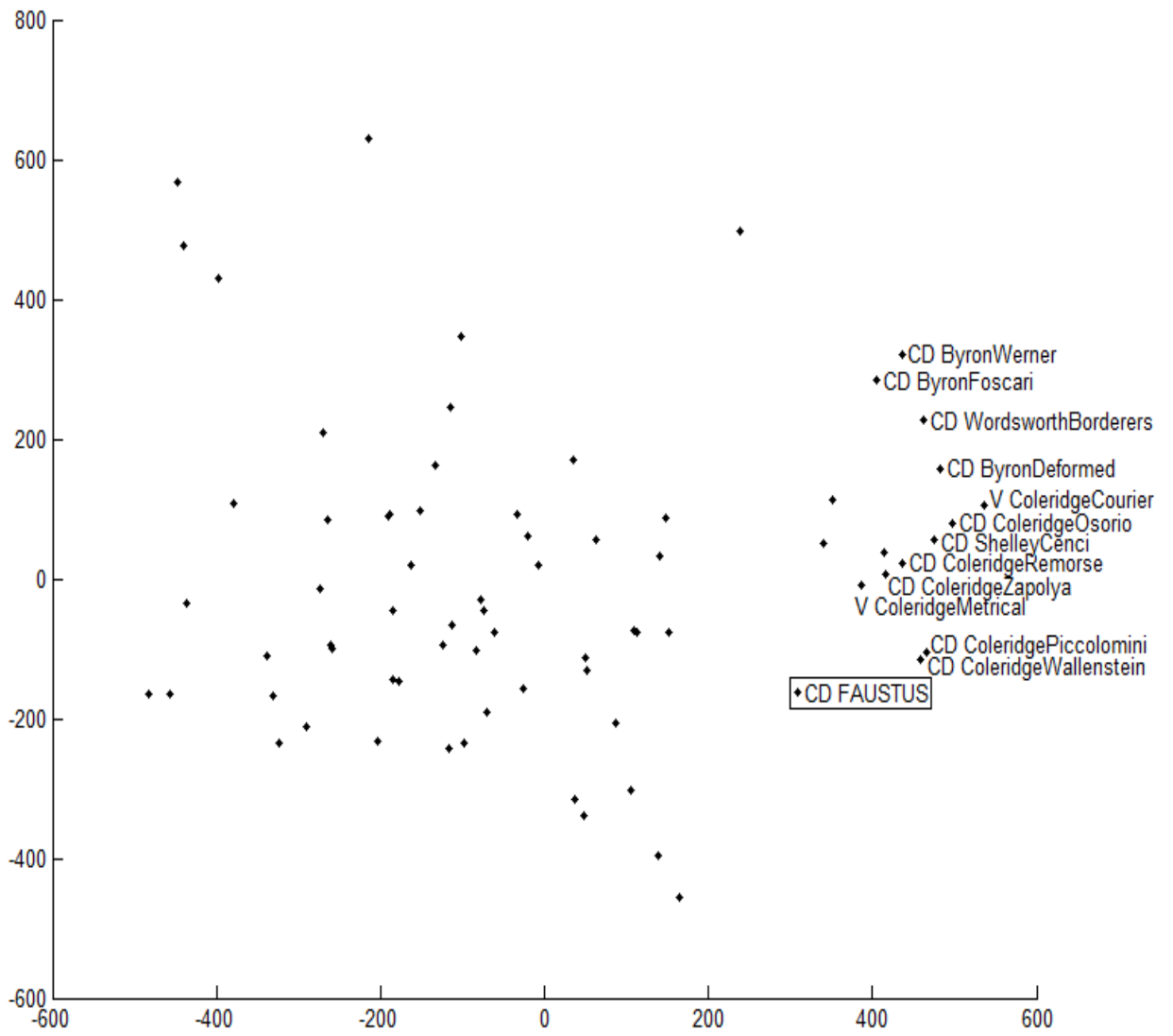


Figure (3.28) Isomap of M380Norm

SOM:

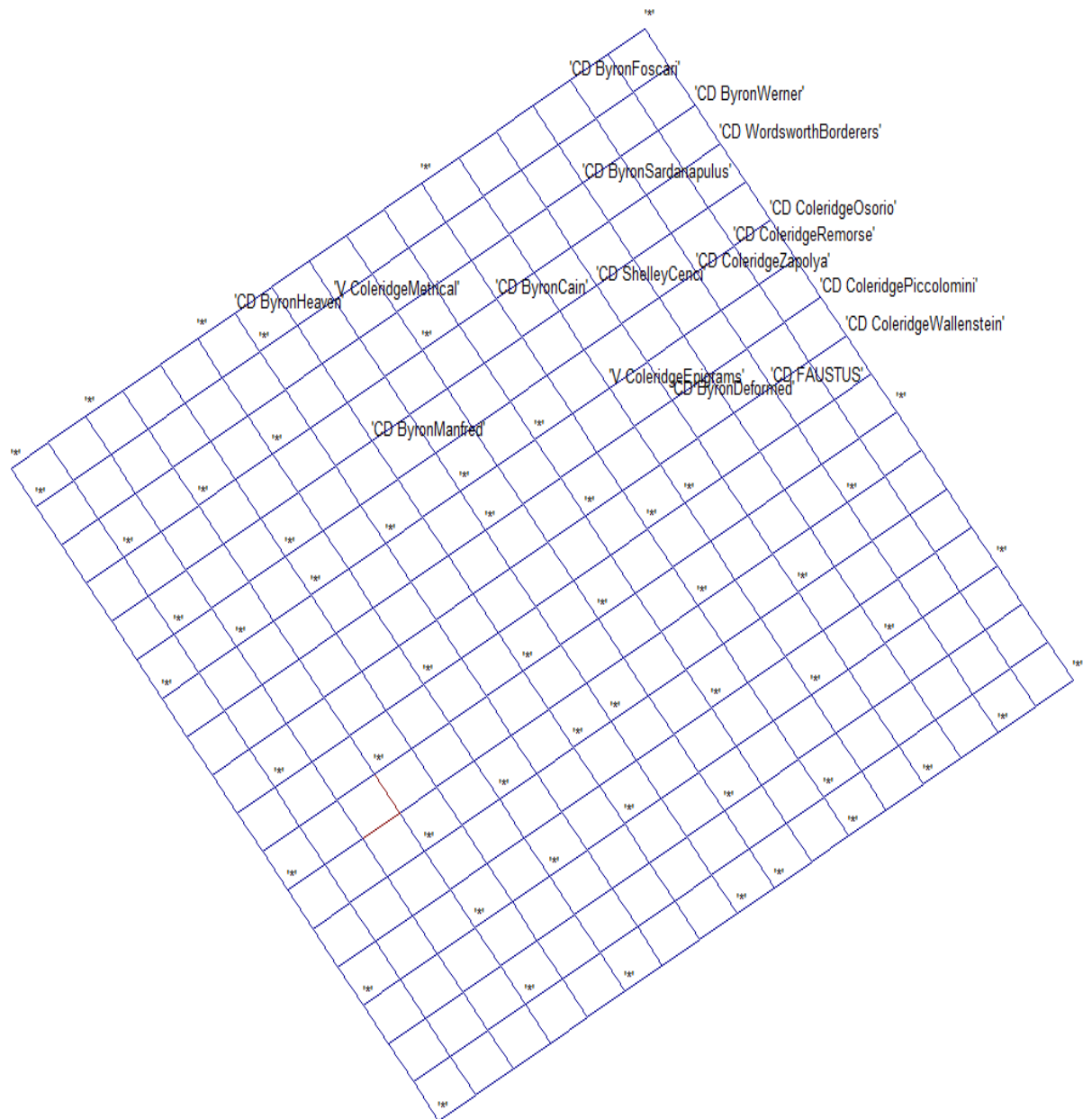


Figure (3.29) SOM of M380Norm

The five clustering methods broadly agree in placing *Faustus* near other closet dramas, and in particular near closet dramas by Coleridge, and even more particularly near *Wallenstein* and *Piccolomini*. A close visual examination indicates a good degree of correspondence among different clustering methods in the way that these texts are placed close to each other in the space by the non-hierarchical methods in a clustering similar to that of the cluster membership that combines these texts all together in one sub-cluster by the average hierarchical clustering. This sub-cluster, however, has clear five sub-clusterings: the first sub-cluster is a one-text cluster, consisting of only *Byron Deformed*;

the second Wordsworth *Borderers*, Coleridge *Remorse*, and Coleridge *Osorio*; the third Shelley *Cenci*, Coleridge *Zapolya*, and Byron *Sardanapalus*; the fourth Byron *Werner* and Byron *Foscari*; the fifth and the final sub-cluster consists of *Faustus*, Coleridge *Wallenstein*, and Coleridge *Piccolomini*. It is obvious that the texts in last sub-cluster are close to each other as if they had been written by a single author. The study calls this sub-cluster a sub-cluster of interest (1), as will be discussed in more detail later in this chapter.

This result confirms rather than falsifies the hypothesis that Coleridge was the author of the 1821 Boosey *Faustus*; the clustering with *Wallenstein* and *Piccolomini* seems particularly significant in that both are translations from German, that is, from plays by Schiller, and, as noted in Chapter One of literature review, Coleridge at the time was a qualified German-English translator of literature who had been asked to translate *Faustus* in 1814 for John Murray and, but arguably, in 1820 for Boosey. This conclusion is of course consistent with the claim advanced by Paul Zall in 1971 and, most recently, with the literary and stylometric pieces of evidence presented by Burwick and McKusick in 2007, as discussed in Chapter One as well.

The main difference between these clustering methods, however, is what a (sub)cluster or neighbourhood of texts are made of. That is, one or two texts that are not assigned to the cluster membership generated by the average hierarchical clustering or not in the neighbourhood of *Faustus* by the non-hierarchical methods are assigned to that (sub)cluster or neighbourhood by one or a couple of these non-hierarchical methods, and vice versa. For example, in PCA and MDS, Coleridge *Cambridge* and CD Byron *Heaven* are placed in the space generated by them. The two methods however differ in that Coleridge *Happiness*, Coleridge *Oldman*, and Coleridge *Metrical* are assigned into the space generated by PCA, while Byron Short poems and Coleridge *Souvenir* are assigned into the space generated by MDS. In Isomap, Coleridge *Epigrams*, CD Byron *Cain*, and CD Byron *Sardanapalus* are not assigned into the space generated by this method.

At this stage of research, the study does not take this similarity as evidence that Samuel Taylor Coleridge is the actual translator of the 1821 *Faustus*. The clustering results just presented suggest no more than that Coleridge is a likely candidate author 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.

3.5 Coleridge and the other translators of *Faustus*:

The results so far support the hypothesis of Coleridge as the author of the 1821 Boosey *Faustus*. Logically, though, they say only that Coleridge is more likely as the author than any of Byron, Shelley, or Wordsworth. But we know that there are other authors who had a demonstrable interest in translating Goethe's *Faustus*, namely Staël, Soane, Anster, Boileau, and Gower, and it is conceivable and one of these might have been the author of Boosey's translation. The final step, therefore, is to add the translations by these authors to the existing corpus, extract function word frequency data from this further-expanded corpus, and then to recluster it to see where in the data space the Boosey *Faustus* sits in relation to the locations of these authors in the space. For the following experiment, therefore, the corpus, therefore, consists of:

- Coleridge's closet dramas: *The Fall of Robespierre* (1794); *Osorio* (1797); *The Death of Wallenstein* (1800); *The Piccolomini* (1800); *Remorse* (1813); and *Zapolya* (1816).
- The closet dramas by Shelley, Byron, and Wordsworth. Shelly's closest dramas: *Faust*; *The Cenci*; and *Prometheus Unbound*. Byron: *Cain: A Mystery*; *Heaven and Earth*; *Manfred: A Dramatic poem*; *Werner; or, The Inheritance*; *The Deformed Transformed*; *The Two Foscari*; *Sardanapalus*. Wordsworth: *The Borderers*.
- The *Faustus* translations by Staël (1813), Anster (1820); Boileau (1820); Gower (1823), and Soane (1821-1825). Here it must be noted that Soane's (1820 and 1821) translations were combined into a single text called Soane 1821.

Because the foregoing results have shown that the Boosey *Faustus* clusters with closet dramas, and because the additional *Faust* translations also belong to this genre, only the closet drama texts are clustered and the verse and prose texts are eliminated. This is done for clarity of presentation.

A function word frequency matrix F4 was generated from the corpus, length-normalized, and dimensionality-reduced to 80 on the basis of figure (3.30).

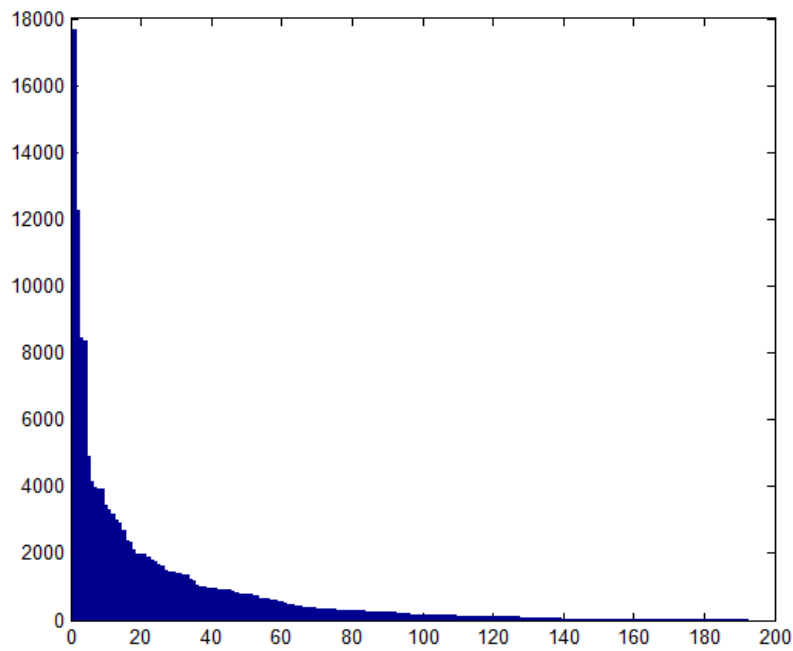


Figure (3.30) The distribution of function words in frequency matrix F4

and the selected 80 highest-frequency function words are shown in Table (3.10):

Word type	Word type	Word type	Word type	Word type	Word type	Word type	Word type
the	may	this	mine	must	you	these	by
my	its	which	nothing	more	but	out	our
with	up	so	without	us	no	nor	or
as	down	at	of	can	what	other	one
from	once	their	that	where	we	himself	she
will	through	shall	his	could	if	in	such
they	within	like	him	most	yet	me	some
then	and	upon	on	till	would	for	those
them	it	should	her	whom	an	all	before
when	he	into	who	to	than	your	though

Table (3.10) The types 80 high-variance function words in figure (3.30)

F4 was cluster analysed using the same methods as before, with the following results.

Single Linkage (Cophenetic correlation coefficient: 8528):

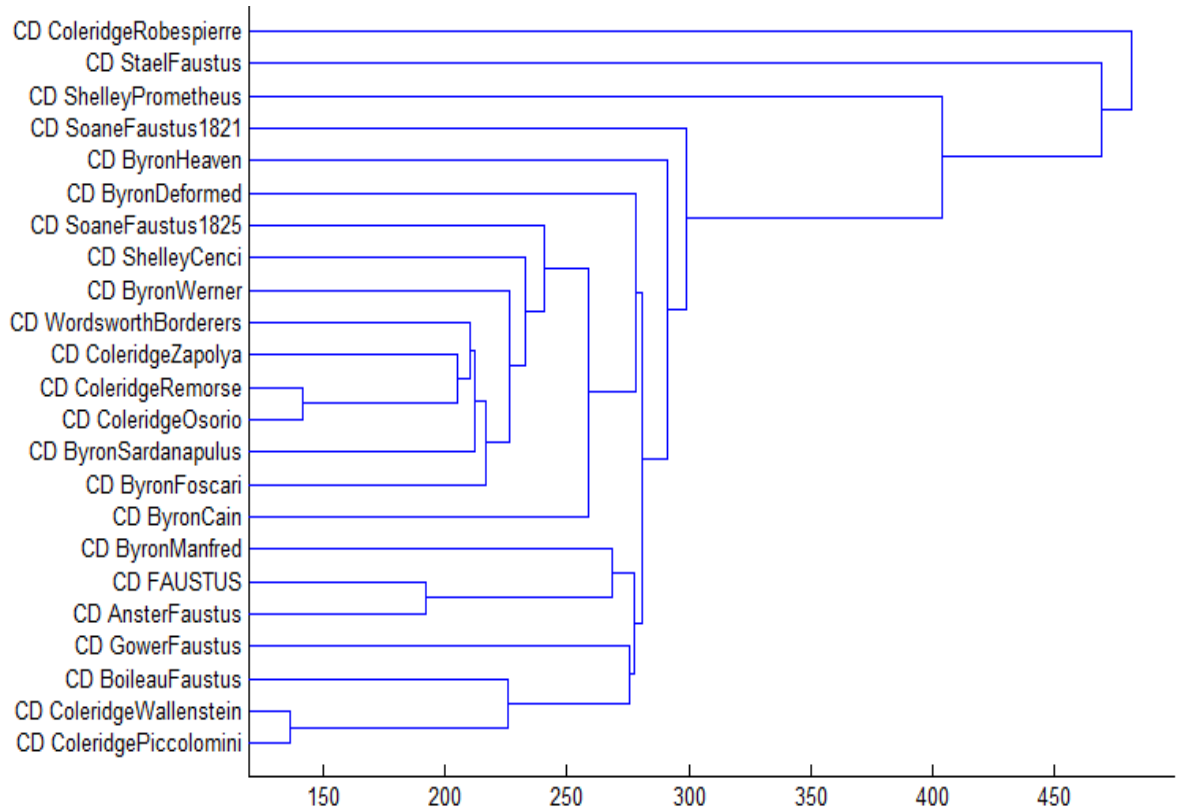


Figure (3.31) Single Linkage. Cophenetic correlation coefficient: 8528

Complete Linkage (Cophenetic correlation coefficient: 0.8729):

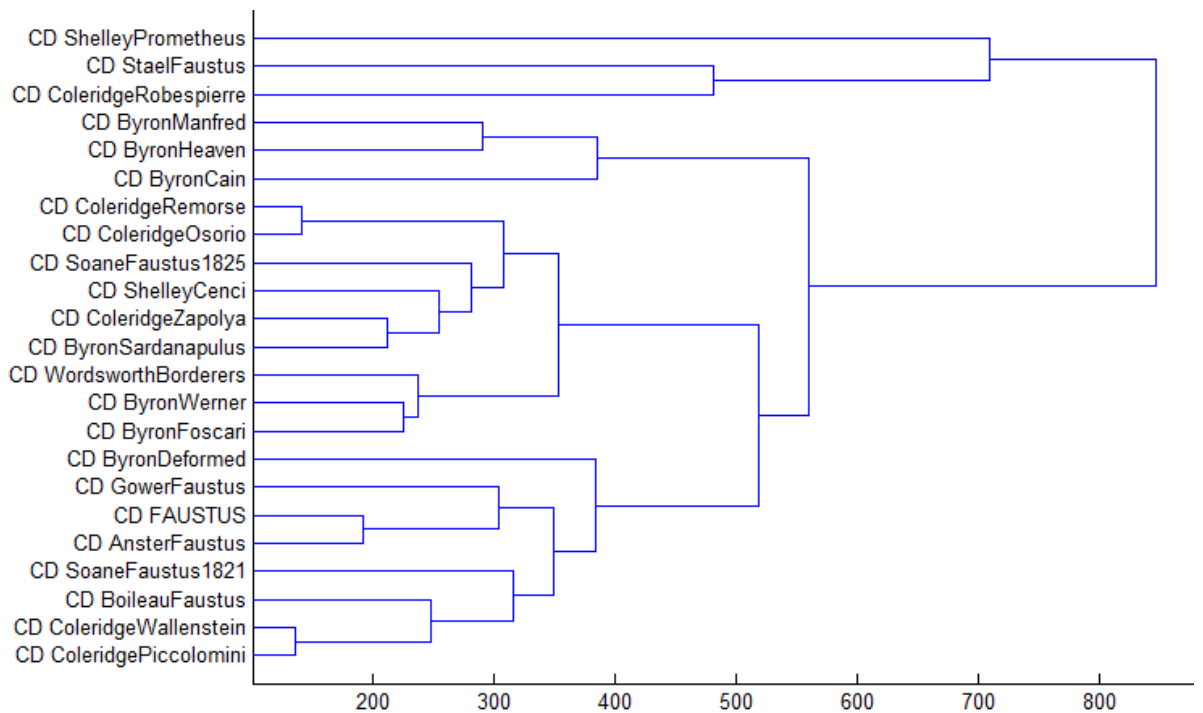


Figure (3.32) Complete Linkage. Cophenetic correlation coefficient: 0.8729

Average Linkage (Cophenetic correlation coefficient: 0.8849):

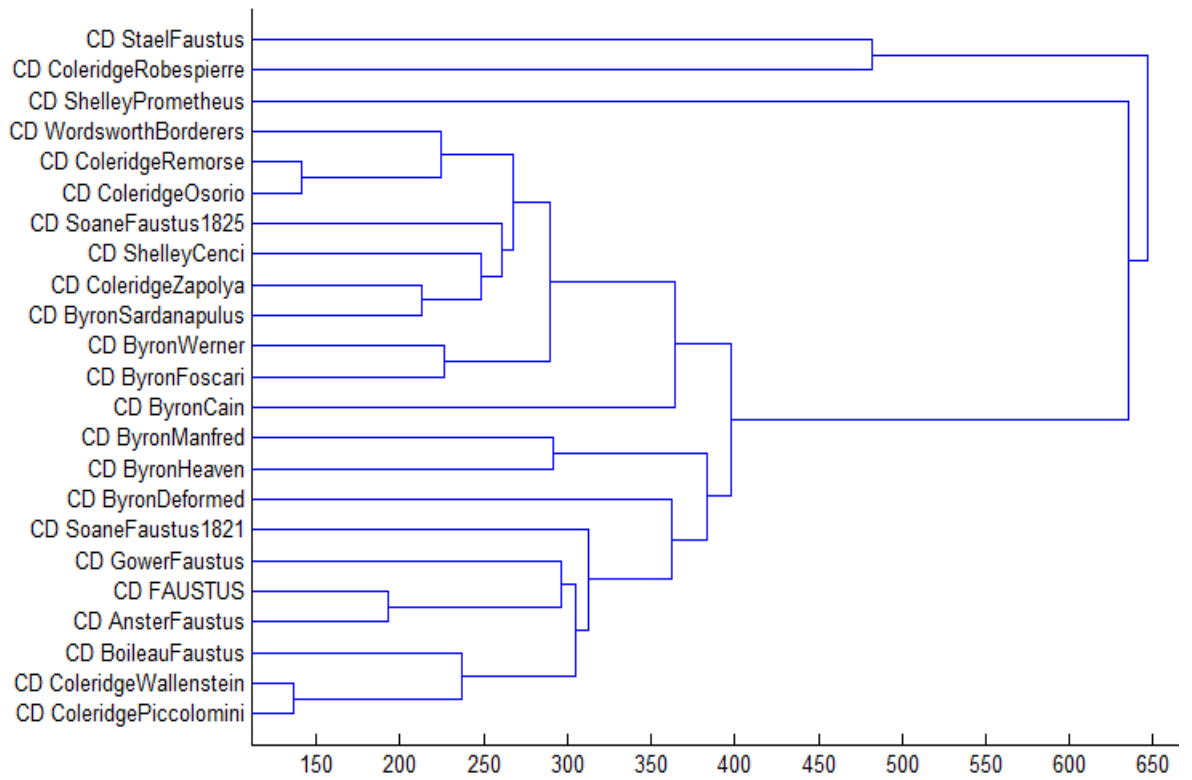


Figure (3.33) Average Linkage. Cophenetic correlation coefficient: 0.8849

Ward linkage (Cophenetic correlation coefficient: 0.4954):

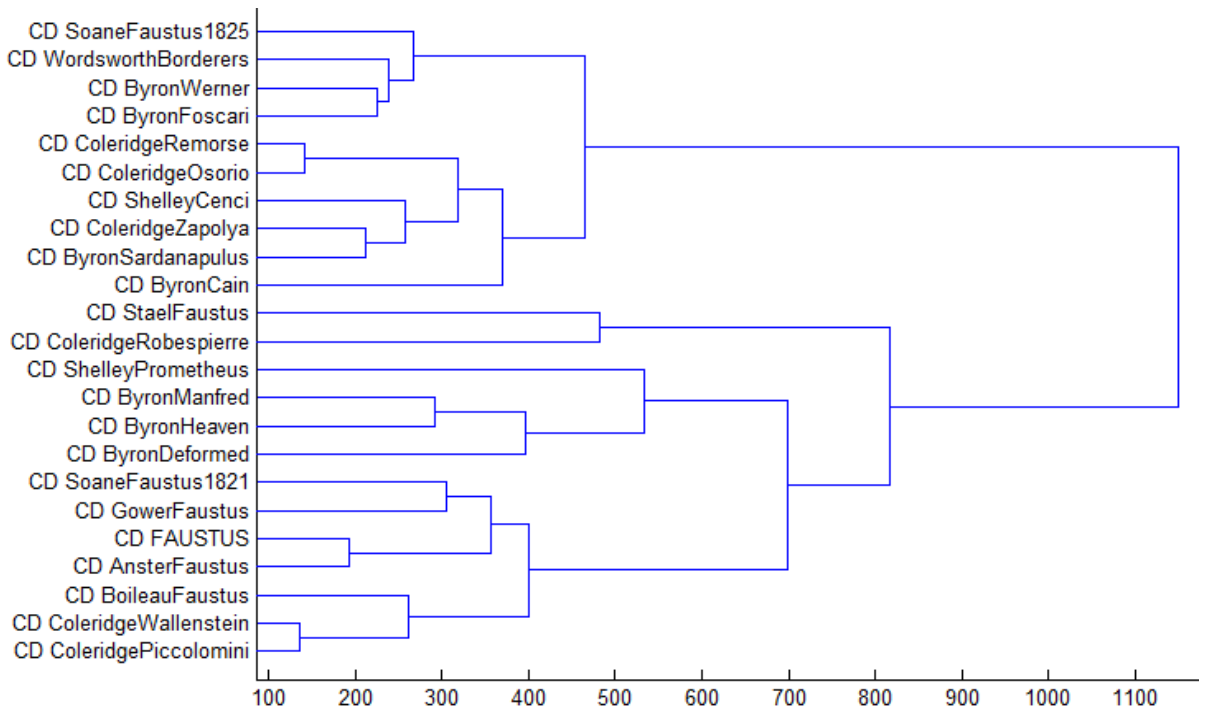


Figure (3.34) Ward linkage. Cophenetic correlation coefficient: 0.4954

The hierarchical clustering tree generated by Average clustering analysis seems to fit M480Norm more better than the clusterings produced by Single, Complete, and Ward analyses.

Hierarchical clustering method	Cophenetic correlation coefficient
Single	0.8528
Complete	0.8729
Average	0.8849
Ward	0.4954

Table (3.11) Cophenetic correlation coefficient for four hierarchical clustering methods applied on M480Norm

PCA:

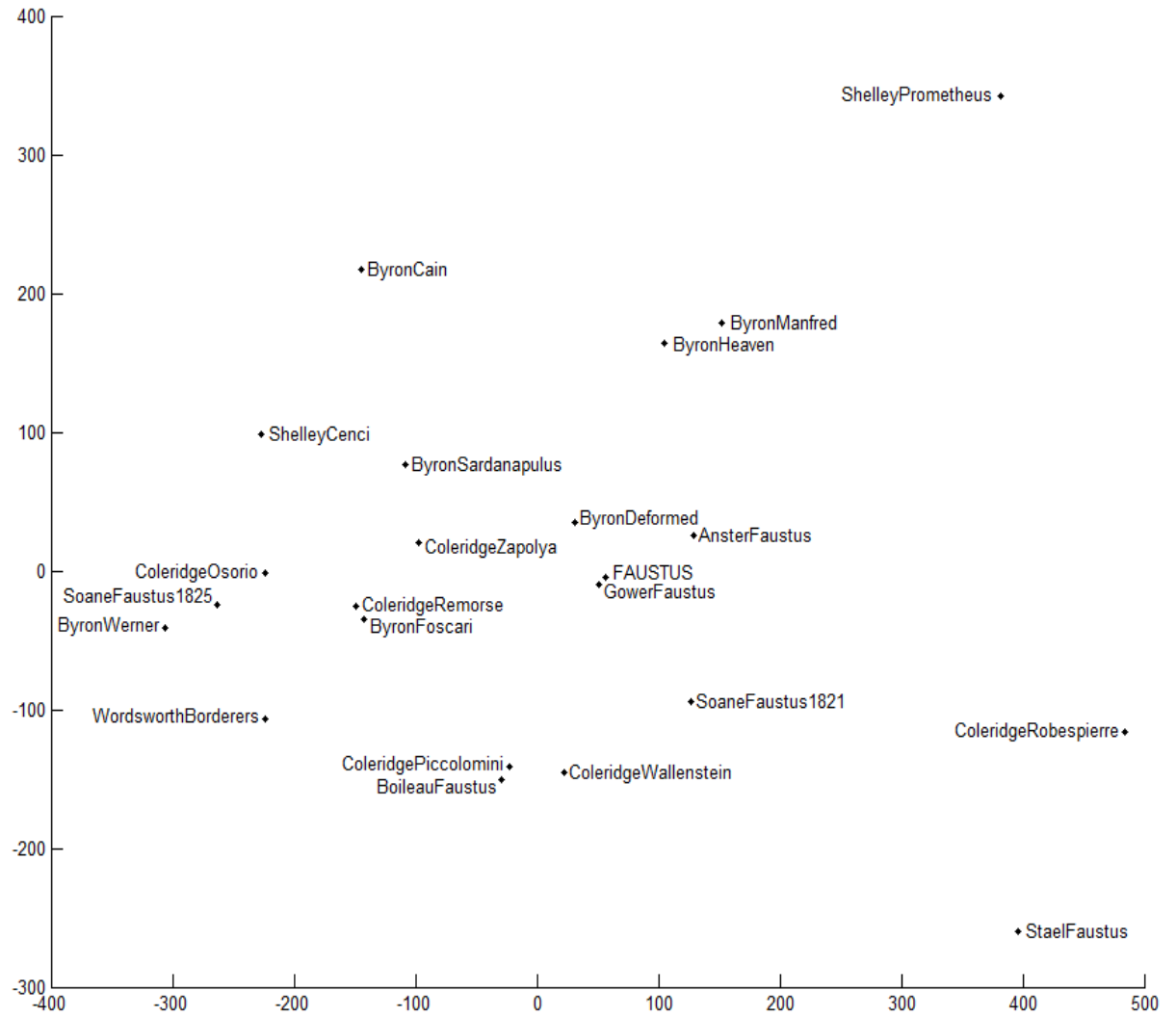


Figure (3.35) PCA of M480Norm

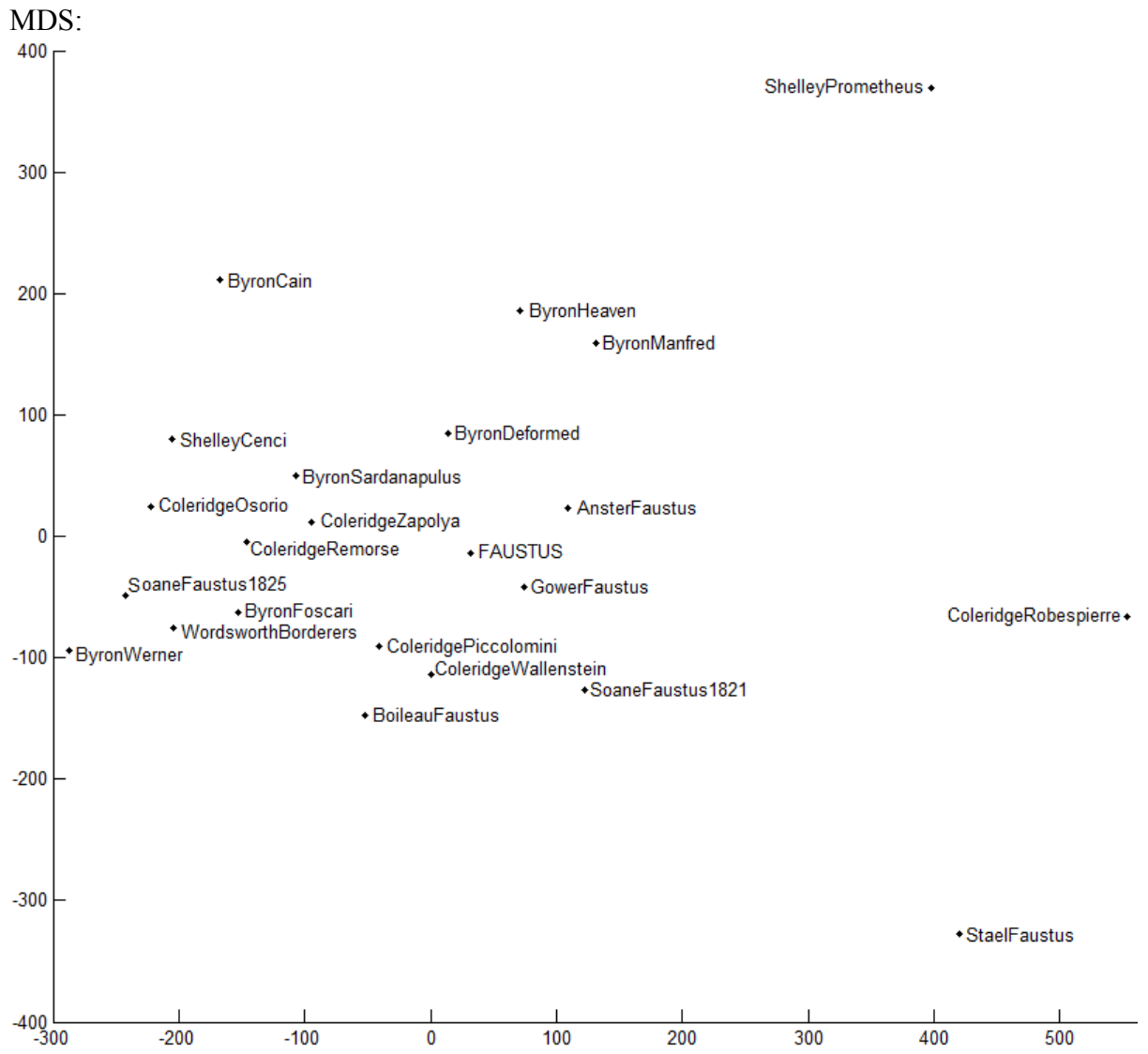


Figure (3.36) MDS of M480Norm

Isomap:

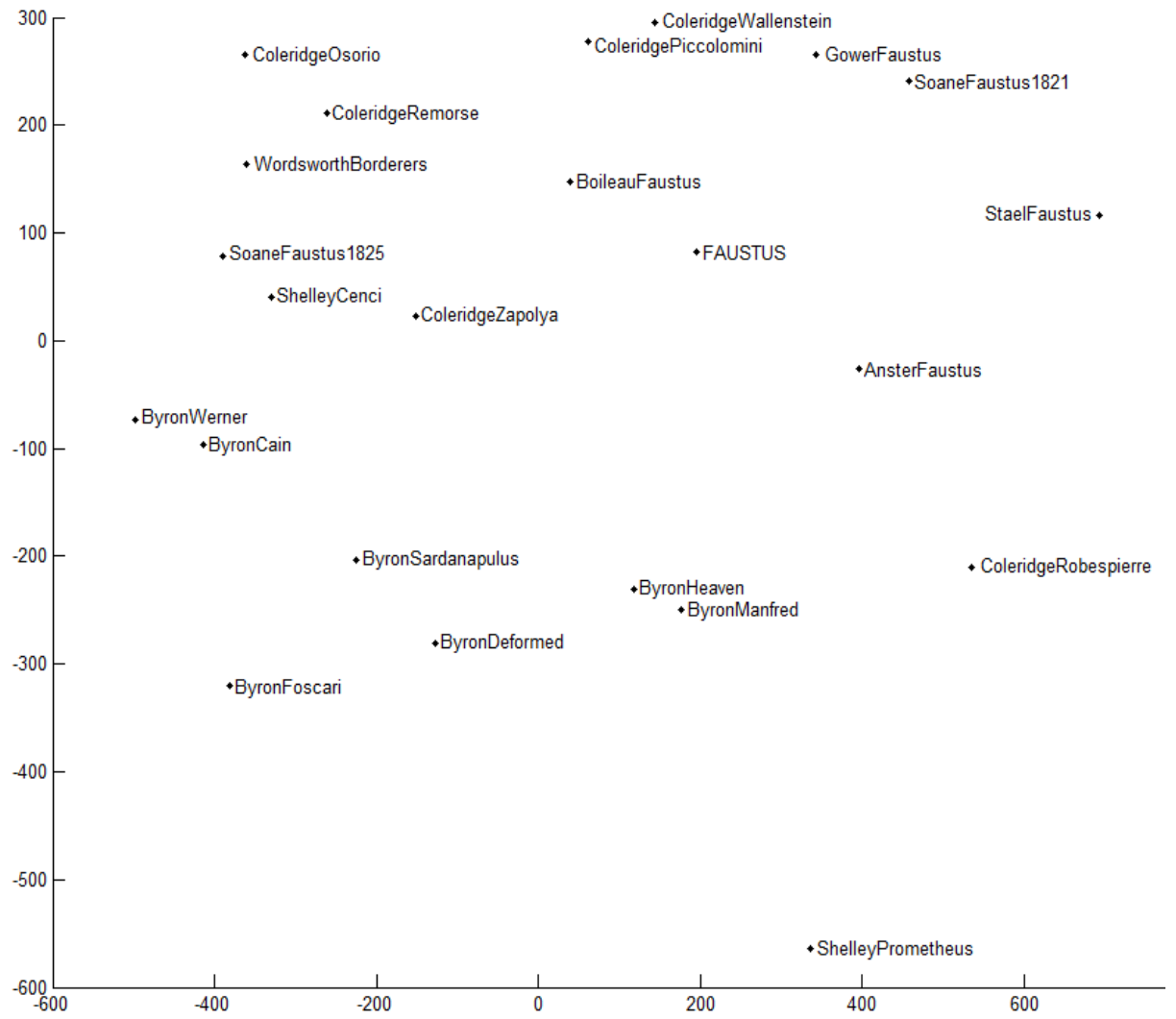


Figure (3.37) Isomap of M480Norm

SOM:

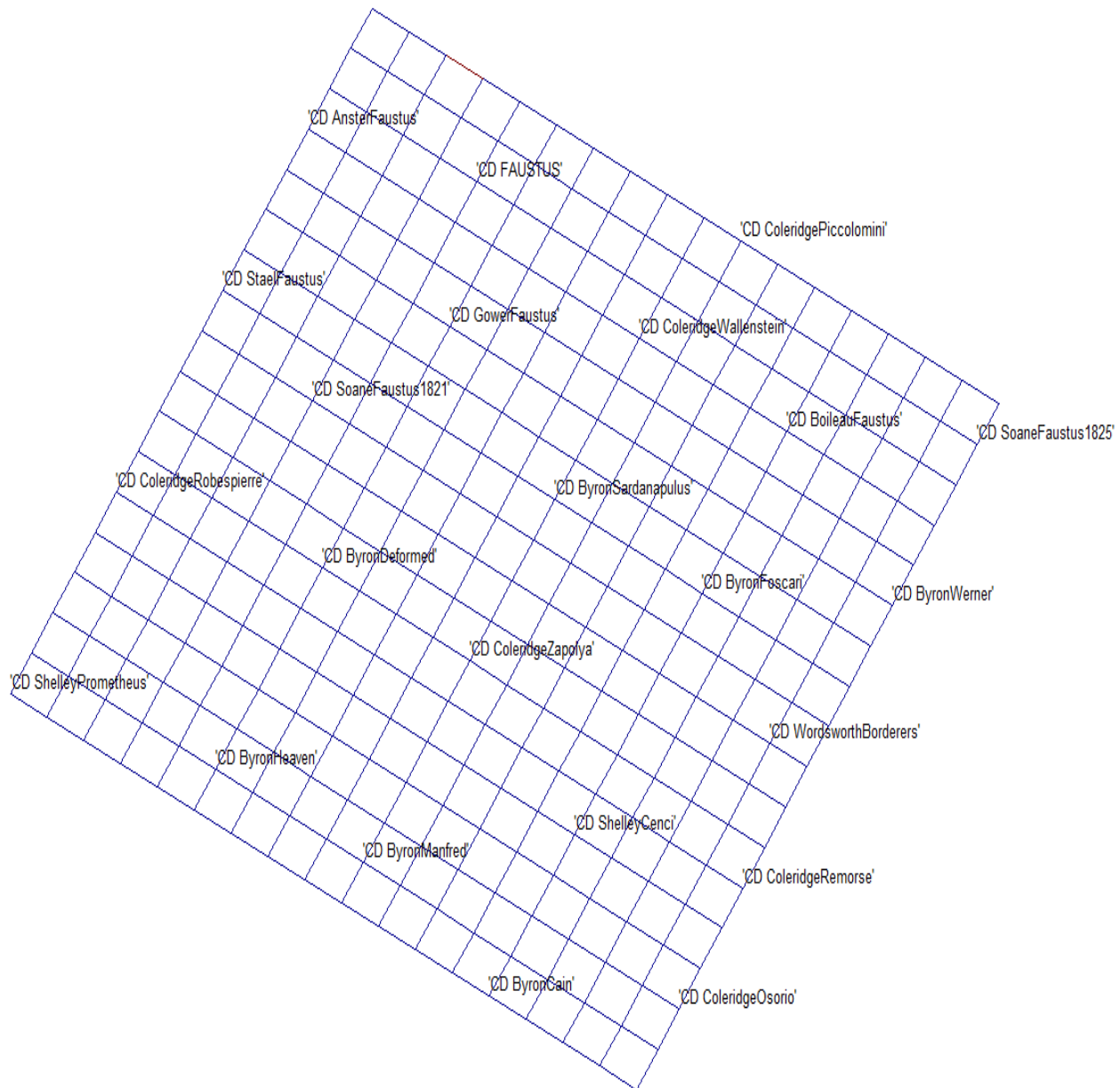


Figure (3.38) SOM of M480Norm

Upon closer examination of all the clustering results, the researcher observes the followings:

- The average hierarchical clustering method groups the closest dramas into three main clusters based on their similarity coefficients or relative similarity from one another. The first cluster consists of CD Stael *Faustus* and CD Coleridge *Robespierre* and the second cluster consists of one cluster representing CD Shelley *Prometheus* on its own. The third cluster comprises two main sub-clusters, each of which is further clustered into small groups of sub-clusters, and more specifically: the first sub-cluster comprises

five sub-clusters: the first consists of CD Wordsworth *Borderers*, CD Coleridge *Remorse*, and CD Coleridge *Osorio*. The second consists of CD Soane *Faustus* 1825 on its own. The third CD Shelley *Cenci* CD Coleridge *Zapolya*, and CD Byron *Sardanapalus*. The fourth CD Byron *Werner* and CD Byron *Foscari*, and the last one consists of CD Byron *Cain* on its own. The second sub-cluster also comprises five sub-clusters: the first consists of CD Byron *Manfred* and CD Byron *Heaven*. The second CD Byron *Deformed* on its own. The third CD Soane *Faustus* 1821 on its own as well. The fourth CD Gower *Faustus*, CD *Faustus*, and CD Anster *Faustus*. The researcher calls this sub-cluster a sub-cluster of interest (2), as will be discussed in more detail in the remainder of this chapter. The fifth and the last sub-cluster consists of CD Boileau *Faustus*, CD Coleridge *Wallenstein* and CD Coleridge *Piccolomini*.

- The Boosey *Faustus* always occurs near the same group of other authors in all the analyses. Based on a very close inspection of the analyses in figures (3.33 and 3.35-8): in the average hierarchical analysis, CD Gower *Faustus*, CD *Faustus*, and CD Anster *Faustus* are placed together in one sub-cluster texts, where, more specifically, CD Gower *Faustus* is clustered with the sub-cluster combining both CD Anster *Faustus* and CD *Faustus*. In PCA, CD *Faustus* is placed close to both CD Gower *Faustus* and CD Anster *Faustus*, but is relatively closer to Gower's. In MDS, CD *Faustus* is placed close to both CD Gower *Faustus* and CD Anster *Faustus*, but again is relatively closer to CD Gower *Faustus* than Anster's. 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 study 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 study.

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. The remainder of the discussion is into parts. The first part one deals with the centroid analysis of authors and the second part with the two sub-cluster texts of interest (1) and (2) mentioned above. That analysis proceeds as follows.

1. From M480Norm, 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 M480Norm 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.
2. The set of individual matrix rows and calculated centroids are co-plotted as bar plots. The relative differences in height of the bar plots indicate differences of usage of the function words corresponding to each of the columns. In other words, here the criterion is only with the amount of variation in the variable centroids or with how much variability is present in a set of bars. 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 sub-clusters of interest because there is much change in the values of that variable throughout text row vectors, i.e. if the difference is large, it is clearly significant.
3. 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 (3.12) and the resulting centroids are then bar plotted onto a bar chart, as shown in figure (3.39):

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

Table (3.12) Function word frequency centroids for 10 authors

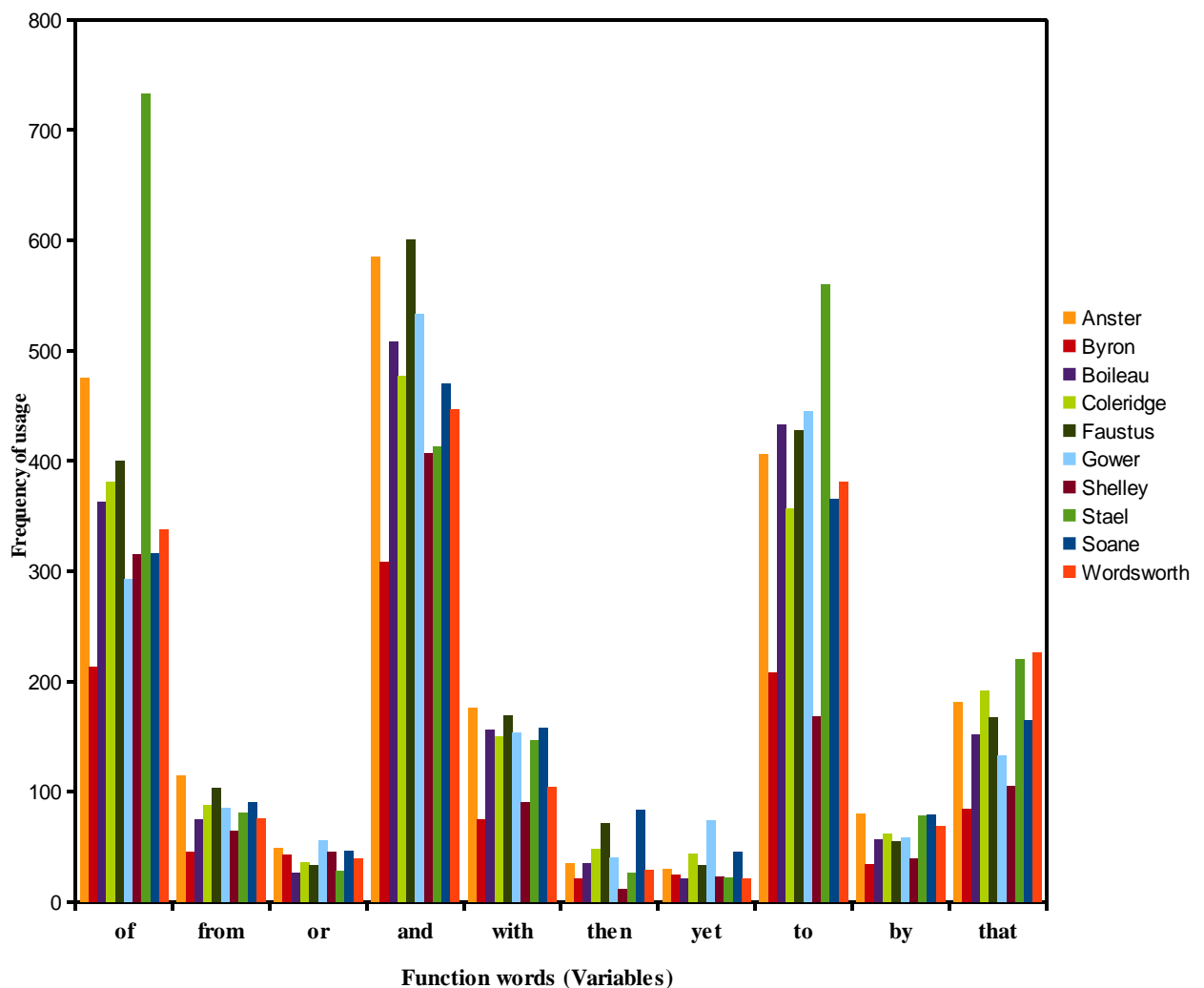


Figure (3.39) Bar plot for 10 authors based on centroid-analysis of 10 FWs

where:

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 (3.12) rather than ordered by size from the smallest to largest or vice versa.

From Table (3.12) and the plot in Figure (3.39), 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 (3.12) above and their corresponding plots in the Figure (3.39), we can draw the following results:

1. 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 Table (3.13):

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

Table (3.13) The amount of variation in the centroids of 10 FWs for 10 authors

2. 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.
3. Coleridge’s usage of this set of 10 function words 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.

Now the study turns to the second part of the centroid analysis which is related to the sub-cluster texts of interest (1) and (2) or the neighbourhood of texts that are clustered all together close to *Faustus* across all the clustering methods. The aim of which is to determine, as above, which one of these selected function words is common or frequent for Coleridge and which is rare or infrequent for all the others.

Based on the forgoing analyses, the two clusters of interest are:

1. Sub-cluster (1) consists of CD Coleridge *Piccolomini*, Coleridge *Wallenstein*, and CD *Faustus*, as in Figures (3.24, 26, 27, 28, 29).
2. Sub-cluster (2) consists of CD Anster *Faustus*, CD *Faustus*, and CD Gower *Faustus*, as in Figures (3.31, 35, 36,37,38).

As we explained above:

- The texts that constituted both sub-cluster texts of interest are collected and saved in a text file document; text file document for each text.
- The centroid for each column in each sub-cluster texts of interest is calculated and saved. The values then represent the centroid characteristics of each function word throughout the texts that constituted a given sub-cluster.
- The centroids are then plotted using bar chart.
- The centroids for each sub-cluster texts of interest can now be compared and interpreted.

The function word centroids of most interest to sub-cluster texts (1) are calculated, as shown in Table (3.14), and the resulting centroids are bar plotted, as shown in figure (3.41):

Word type	Faustus	Piccolomini	Wallenstein
of	400	728	500
from	103	186	163
or	33	59	41
and	601	1022	850
with	169	356	205
then	71	99	70
yet	33	83	43
to	428	950	700
by	55	97	69
that	167	474	300

Table (3.14) Function word frequency centroid for sub-cluster texts of interest (1) based on 10 FWs

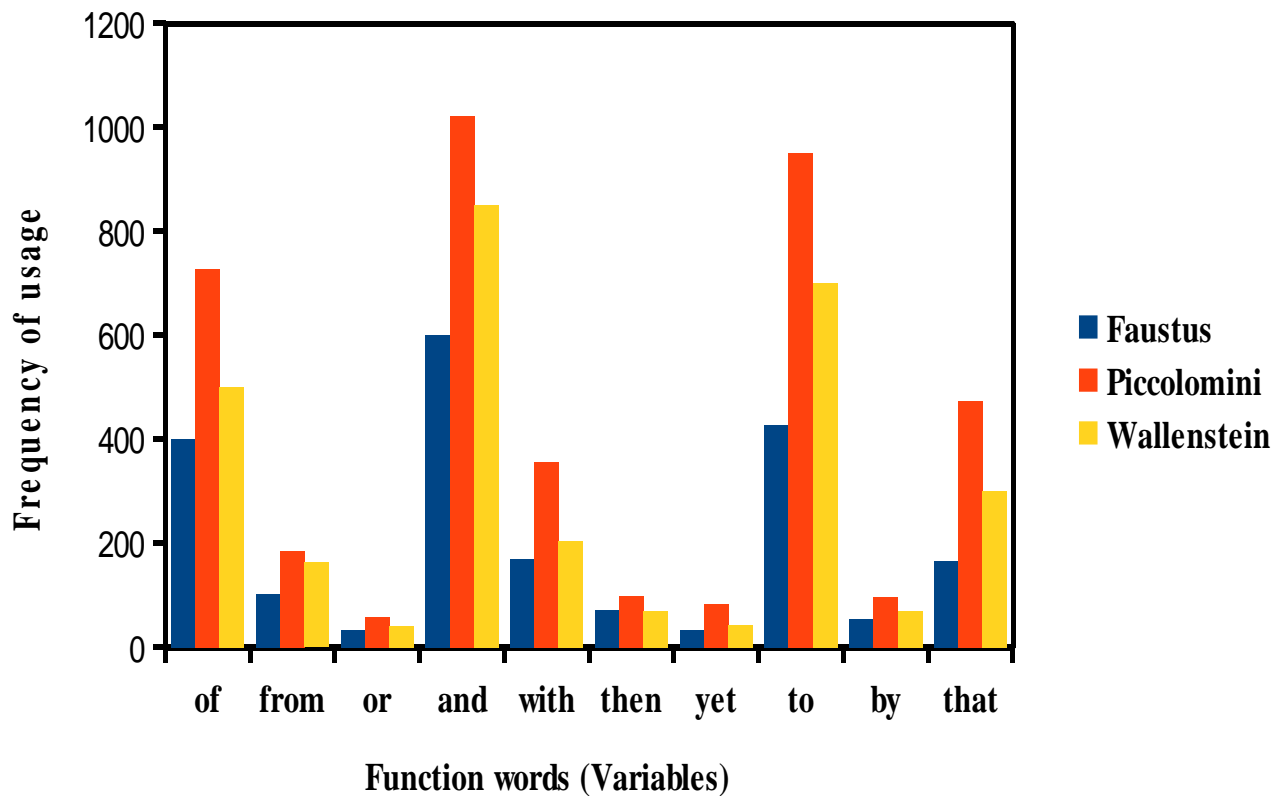


Figure (3.40) The usage of 10 high variance function words across *Faustus*, *Piccolomini*, and *Wallenstein*

It can be seen from the plot in this figure, first of, that there is relatively much less variation in the usage of ‘of’, ‘from’, ‘or’, ‘and’, ‘with’, ‘then’, ‘yet’, and ‘by’ across Coleridge’s two plays as represented by the amount of variation in each bar. Secondly, Coleridge’s usage of ‘and’, ‘to’, and ‘that’ is very different from that of the 1821 anonymous translator of *Faustus*.

The overall indication therefore is that there are differences between Coleridge’s two dramas and the 1821 anonymous translator of *Faustus* and that the function words ‘of’, ‘to’, ‘that’, and ‘and’ are the main determinants for these differences based on the amount of variation in their corresponding centroids shown in Table (3.14) above, which are calculated and shown in Table (3.15) below:

Word type	Amount of variation
of	28261.2133
from	1836.3024
or	177.3113
and	44804.0333
with	9844.1132
then	271
yet	700
to	68161.3443
by	457.3333
that	23702.3453

Table (3.15) The amount of variation in the centroids of 10 function words for *Faustus*, *Piccolomini*, and *Wallenstein*

The function word centroids of most interest to the sub-cluster texts (2) are calculated, as and shown in Table (3.16), and the resulting centroids are bar plotted, as shown in figure (3.41):

Word type	Faustus	Anster	Gower
of	400	475	293
from	103	115	85
or	33	49	56
and	601	585	533
with	169	176	154
then	71	35	40
yet	33	30	74
to	428	406	445
by	55	80	58
that	167	181	133

Table (3.16) Function word frequency centroid for sub-cluster texts of interest (2) based on 10 function words

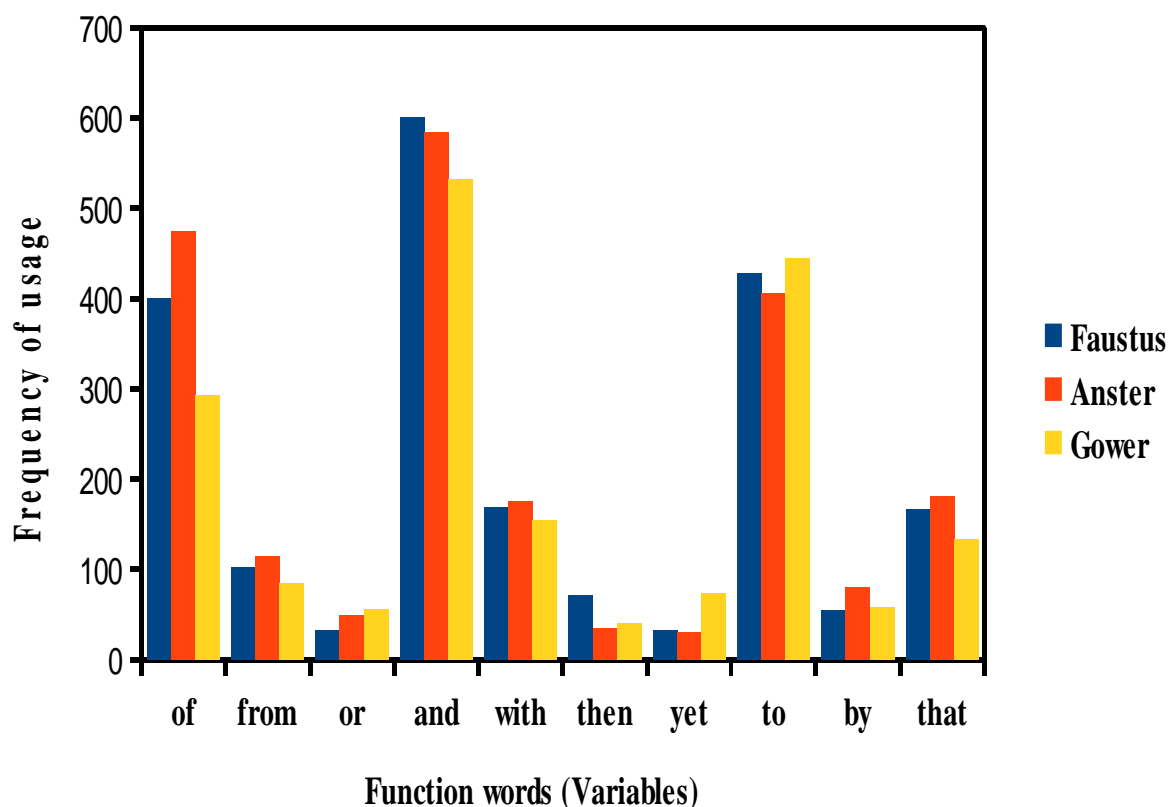


Figure (3.41) The usage of 10 high variance function words across *Faustus*, Anster, and Gower

In the plot of Figure (3.41), both authors Anster and the 1821 anonymous translator of *Faustus* use ‘from’, ‘and’, ‘with’, ‘yet’, and ‘that’ with almost similar frequency as represented by the height of the bars, but Gower’s usage of these words is different. Gower and the 1821 anonymous translator of *Faustus* use ‘to’, ‘by’, and ‘then’ with almost similar frequency, but Anster’s usage of these words is different.

Based on the amount of variation calculated for the centroid bars in Table (3.16) and shown in Table (3.17):

Word type	Amount of variation
Of	8366.3333
From	228
Or	139
And	1264
With	126.2654
Then	380.3065
Yet	604.1327
To	382.3054

By	186.0032
That	609.4109

Table (3.17) The amount of variation in the centroids of 10 function words for *Faustus*, Anster, and Gower

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 chapter will justify this claim.

Chapter Four

Additional Interpretation

In the previous chapter we carried out the analysis using different clustering methods and noted Anster and Gower were closer to Boosey than to Coleridge or any other of the authors tested. The final result was that the anonymous 1821 *Faustus* translation and the two other translations of the play by Anster and Gower were mathematically clustered together since they share the most distinctive function words, and an additional interpretation was promised on this result. This chapter provides that 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. This result seems to be identical with the result of the word length analysis in McKusick's stylometric section when he compared the 1821 *Faustus* with the two translations by Anster and Gower. In these two graphs and comparisons, which were reproduced and cited in figures (1.2) and (1.3) in Chapter One, and according to the researcher's direct observation, *Faustus* appears to be relatively more similar to Anster translation of *Faustus* than to Gower's. This result also seems to be identical with Hugh Craig (2008: 87) who believes that "it seems unlikely that a writer who had produced one translation would then produce a second quite different one".

In such a case, the question is: can the anonymous 1821 *Faustus* be attributed to Anster or should it rather be attributed to Gower based on this new evidence? The answer is no; the remainder of this chapter justifies that claim. The argument will be 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. The importance of understanding and explaining these two issues cannot be ignored or passed over it silently since this has the effect of clustering the translations by Anster, the 1821 anonymous translator, and Gower all together. In the discussion that follows, familiarity with Section 2.2.3.3, Points (3) and (4) 'the use of function words' in general and with Chapter 1, 'Literature Review' is assumed.

On the face of it, the language of Goethe's *Faustus* in its diction, syntax, and tone would appear to be difficult to translate, even in prose, into English. There are many issues involved in translating any piece of literature from German to English in general and also there are specific difficulties that are encountered in translating *Faustus* in particular (e.g. Constantine, 2006; Classe 2000; Hauhart: 1909; Haney, 1902; Taylor, 1856; Anster, 1835). More specifically, there are many linguistic and lexical difficulties or difficult choices (e.g. word choices, syntactic constructions, feminine rimes, meters, rhythm, etc.) that had to be made in producing the translation of *Faustus* into English. It is, however, possible to say that even the choice of some function words which are used to signal the structural relationships and hold words to each other within the clauses or sentences presented a problem for the translators, which needed to be considered right from the start of the translation. This question would be possible, but is open to argument. What makes translating such words into English difficult is the fact that in English the same function word can be a preposition at times and something else at others (adverb, conjunctions, etc.): what matters here is the function a word has in the given situation. In German, however, this property is a bit different: there are different words for different functions. Take the preposition "von" for example:

Er ist enttäuscht **von** dir

He is disappointed **in** you

That does not mean that "**von**" means "in". Actually "**von**" does not mean "**in**" most of the time, only in this context it does. Consider these two examples:

Wann kommst du **von** der Arbeit zurück?

When are you getting back **from** work?

Ich komme **von** England

I come **from** England

German also, as a special feature, has many one word prepositions, as in:

Ich parkte mein Auto direkt **vor** der Hochschule

I parked my car **in front of** the university

and:

Er fand ihn in einem Stuhl **neben** dem Bettzimmer entspannen

He found him relaxing in a chair **next to** the bed room

Note that **vor** is different in this sentence:

Maria ist **vor** einer Stunde in die Bibliothek gegangen

Maria went to the library an hour **ago**

Prepositions are frequently used in older and modern texts and are often a source of errors and misunderstanding for translators. According to Volk (2006: 84-6), frequent or usual German prepositions are monomorphemic words. Many of the less frequent prepositions are complex (or polymorphemic) since they are derived from other parts of speech such as nouns (e.g. *angesichts*, *zwecks*), adjectives (e.g. *fern*, *unweit*), participle forms of verbs (e.g. *entsprechend*, *während*, *ungeachtet*), or lexicalized prepositional phrases (e.g. *anhand*, *aufgrund*, *zugunsten*). The first source of problem is that only monomorphemic prepositions constitute prepositional objects, pronominal adverbs, and prepositional reciprocal pronouns and this process requires different grammatical case requirements. Monomorphemic prepositions such as *durch* (through), *für* (for), *gegen* (against), *ohne* (without), *um* (about) are governed by accusative case by taking a direct object, monomorphemic prepositions such as *aus* (from, out of), *bei* (at, near), *mit* (with, by), *nach* (after, to), *von* (by, from), *zu* (at, to) are governed by dative case by taking an object, and monomorphemic prepositions such as *an* (at, on, to), *auf* (at, to, on, upon), *hinter* (behind), *in* (in, into), *neben* (beside, near, next to), *über* (about, above, across, over), *unter* (under, among), *vor* (in front of, before, ago), *zwischen* (between) are governed by both accusative and dative case. On the other hand, most of the derived prepositions such as *angesichts* (in view of, in the face of), *bezüglich* (relative to, relating to), *dank* (thanks to) are governed by genitive case by taking an object in the genitive case. There are only a few common genitive prepositions in German and most of the time the genitive prepositions can be translated with "of" in English. One set of these prepositions, such as *während* (during, while, whereas, in the course of), is in the process of changing from genitive case to dative case. Another set, such as *je* (ever, at any time, always, at all times) and *pro* (per), does not show overt case requirements and is used with determinerless noun phrase. Next set of these prepositions, such as *bis* (until, to, by), either takes another preposition (*in*, *um*, *zu*) or connects with the particle *hin* (down, away) in addition to a preposition.

The second source of difficulty is that all frequent prepositions (e.g. *an*, *für*, *in*, *mit*, *zwischen*) have some homegraphic functions which need to be clearly marked by their position within the clause. The most frequent homographic functions are:

-separable verb prefix. For example *ab* (from...onwards), *auf*, *mit*, *zu*, *an*, as in: (Wann haben sie **angefangen**?) means (When did they begin?)

-clause conjunction. For example *bis* and *um*, as in: (**bis** vor 3 Wochen hatte nie ein Wort Deutsch gelernt und jetzt spreche ich fließend) means (**until** 3 weeks ago never a word of German had learned and now I speak fluently).

-adverb. For example *auf*, *für*, and *über* often in idiomatic expressions, as in: (**auf** und davon) means (gone) or (**über** und **über**) means (all over).

- infinitive marker. For example *zu*, as in: (Ich mag es Französisch **zu** sprechen) means (I like **to** speak French).

-proper name component. For example *von*, as in: (Sie hat keine Zeit **zu** lesen) means (She has no time **to** read).

-predicative adjective. For example *an*, *auf*, *aus*, *in*, *zu*, as in: (die Maschine ist **an/aus**) means (the machine is **on/off**) or (die Tür ist **auf/zu**) means (the door is **open/closed**).

(Bauer, 2015a; 2015b; 2015c; Hatherall and Hatherall, 1995)

On function words in German see, for example, Jones and Tschirner (2006), Dikken and Tortora (2005), Schnorr and Forst (1995). Detailed discussions of syntax and semantics of prepositions or function words in German can be found, for example, in Dewell (2015), Volk (2006), Hatherall and Hatherall (1995).

Given this conundrum, many function words could not be translated adequately into English: There is no one-to-one equivalence between the function words in German and another function words in English. German typically just has one word preposition (i.e. monomorphemic) for each situation while English uses combinations of words. With the confusion this can cause, the researcher believes that the text of *Faustus* was not easy to translate without making considerable changes to it, syntactic and/or lexical borrowing, or, to say the least, a word-for-word German English translation and imitation.

As we have already noted in Chapter One, six English translations of parts of Goethe's *Faustus* (Part I) was made between 1813 and 1823. Such translations were generally in the form of prose or as blank verses mixed with prose summaries. Some translations covered large selections of the play while others had a more narrow focus. Nearly all the

translators who attempted to translate the play either ignored some of the passages because they found no exact words equivalent in English or simply imitated the Goethe's varied meters and verses as closely as possible. And because of this the translations for some pieces were hardly English poetry. In this connection the question of the qualifications of an ideal translator or the translator's knowledge may be brought up to learn something about the qualifications of Anster and Gower for the task. There is a limited number of reviews devoted to the qualifications of these two authors for the task, the following is only a brief account on this question, and is based on Hauhart (1909: 99-103, 121-4) and Burwick and McKusick (2007: 223-227, 280-2).

John Martin Anster was an English poet and translator. He was born in the year 1793 in the country Limerick, Ireland. He was educated Trinity College, Dublin. His contribution to poetry began in 1815 when he was only twenty two and published *Ode to Fancy*, some sonnets, and two poems in imitation of Coleridge's 'A Poet's Haunt' and 'Solitude: An Ode'. Four years later, in around 1819, he won a prize at Trinity for his 'Lines on the Death of Princess Charlotte' which was published in his second poetic volume in *Blackwood's Magazine* as *Poems: With Some Translations from the German*. About this time Anster began his translation of fragments of Goethe's *Faust* into English. It first appeared in *Blackwood's Magazine* in 1820 as *Goethe's Faust*. In this edition, Anster explained his strategy of translating parts from the Goethe's play by saying that "To verbal fidelity, I can, of course, make no claim; yet I have not wilfully deviated from it. I have not sought to represent my author's thought by 'equivalents', ...I should say that I always have given a perfectly accurate translation of the very words, now and then expanding the thought by the addition of a clause, which does little more than express something more fully implied in the German than in such English phrases as occurred to me." However, Goethe praised Anster for the faithfulness of the translation to the original as well as for the quality of English poetry into which it had been turned. Anster later completed his translation of the first part of the play in 1835. This translation appeared in book-form as *Faust: A Dramatic Mystery; The Bride of Corinth; The First Walpurgis Night*. In the preface to this edition, Anster admitted that "There are peculiarities both in the conception and in the structure of the drama which seem to require a few words of notice. The easiest and least formal manner of discussing the subject is to state the difficulties with which those peculiarities embarrass a translator". The second part of this completed translation published also in book-form in 1864 and appeared as *Faustus: The Second Part, from the German of Goethe*.

Besides *Faust*, Anster translated many poems from the French and German and contributed for many years to *Blackwood's Magazine*, the *Dublin University Magazine*, *The North British Review*, etc. His best-known works include a third collection of poems which was published in 1837 and appeared as *Xeniola*. This collection included translations of 'Ranz des vaches', from Friedrich Schiller's *William Tell*, scenes from Friedrich de La Motte Fouque's *The Pilgrimage*, S. E. W. von Sassen's 'Memory', Dallwitz's *The Five Oaks*, and Karl Theodor Korner's *Gipsey Song*.

As for Gower's literary career, he was born in 1800 in London and educated at Eton and Christ Church, Oxford. Gower started to write poetry at an early age or before he was twenty. He published several poems for sole use, which followed up after a short time by the publication of some translations of German lyrics and a few original poems. His first translation of extracts from Goethe's *Faust* was published in 1823 in book-form entitled *Faust: A Drama in Verse by Goethe, and Schiller's song of the Bell*. For almost ten years this version was the worst English translation of *Faust* in existence. In his introduction Gower revealed that "he left sundry passages unattempted where he was convinced of his own inability to transfer their spirit to a translation...consideration of decency had also in a few instances prevented him from proceeding" and that "that the passages in question were not indispensable for the understanding of the story. Of the Prolog he gave only the Archangels' Chants, omitting the rest of the scene, and appending a note in which he briefly gave the contents of the dialogue between the Lord and Mephistopheles, stating that he omitted it in the translation, because the "Tone of familiarity on both sides is revolting in a sacred subject". Goethe's himself showed his disapproval of omitting the Prologue in Heaven or anything besides it, saying that "How so, that is quite unobjectionable, the idea is in Job. He did not perceive that that was the aggravation and not the excuse." In fact, Gower possessed some poetic ability, but his comprehension of German was entirely inadequate for a good translation of *Faust*. Some even said that he did the task as a practice while learning the German language. Nevertheless, Gower published a second edition in 1825. This translation appeared as *Faust, A Drama by Goethe, with other translations from the German*. Gower also translated *Wallenstein's Camp* in 1830, and prepared to publish his translation of Victor Hugo's *Hernani*. In 1839 he visited the Mediterranean and the Holy Land. His impressions of travel were recorded in *Mediterranean Sketches* (1843) and in the notes to a poem entitled *The Pilgrimage*.

It is obvious that the translations of the play by Anster and Gower had portions of considerable merit and poetical insight which vary considerably from one translation to

the next. Given the difficulties of the German text, Anster obviously took such cases into account and translated the greater part of Goethe's verses in blank verse. He tried to avoid the really difficult features of a good translation and to follow the meaning of the original as closely as possible and as is consistent with the nature of the English language. After all, many of the reviewers focussed on Anster's translation of *Faust* and agreed on the "ease", "grace", and "fluency" of his verses in some passages. Yet not everyone agreed that his translation was a satisfactory representation of Goethe's text. Therefore, Anster was seen as too much of a poet to be a close translator, and the translation itself was described, for example, as "an almost incredible dilution of the original" and "a brilliant paraphrase".

With all the difficulties confronting Gower on account of his lack of knowledge in the German language, some of the reviewers were willing to admit that his verses had good quality in some instances and that when he understood the meaning, he often produced a very good translation. Of course Gower's translation had passages of good translation, but for other reviewers a whole translation "must be condemned".

All things considered, one has to point out that translating the words of the original text of *Faust* slides over into borrowing from one author into another. A few examples will suffice 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.

Line number	Anster 1820	Anonymous 1821	Gower 1823
354-364	Alas! I have explored Philosophy, and law, and medicine, And over deep divinity have pored, Studying with ardent and laborious zeal And here I am at last , a very foal,	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:	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,

	<p>With useless learning cursed, No wiser than at first! They call me doctor— and I lead These ten years past my pupils' creed,</p>	<p>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</p>	<p>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,</p>
410-423	<p>Where even Heaven's light so beautiful Thro' the stained glass comes thick and dull— 'Mong volumes heaped from floor to ceiling, Thro' whose pages worms are stealing— Dreary walls— where dusty paper Bears deep stains of smoky vapour— Glasses— instruments— all lumber Of this kind the place encumber— All a man of learning gathers— All bequeathed me by my fathers— Are in strange confusion hurled! Here, Faustus, is thy world— a world! And dost thou ask, why in thy breast The fearful heart is not at</p>	<p>Where thro' the painted glass ev'n heav'n's free light Comes marr'd and sullied, narrow'd by dark heaps Of mould'ring volumes, where the blind worm revels— Of smoke-stain'd papers, pil'd ev'n to the roof— Glasses and boxes— instruments of science— And all the old hereditary lumber Which crowds this cheerless chamber. This is then Thy world, O Faustus! this is called a world! And dost thou ask, why thus tumultuously Thy heart is throbbing in thy bosom why Some nameless feeling</p>	<p>And ask I why my heaving heart Is beating in its sullen madness? And ask I why the secret smart Has dried the spring of life and gladness 'Tis that instead of air and skies, Of nature's animated plan, Round me, in grinning ranks, arise The bony forms of beast and man, Wake then, my soul, thy wings expand: This book by Nostradamus' hand, Sigil and sign shall make thee fly Uncheck'd, unwearied, through the sky. Wake then, my soul! the signs of power</p>

	rest! Why painful feelings, undefined, With icy pressure load thy mind!	tortures ev'ry nerve, And shakes thy soul within? Thou hast abjur'd	Point to the destined tide and hour
428-435	Ha! what new life divine, intense, Floods in a moment every sense; I feel the dawn of youth again, Visiting each glowing, vein! Was it a God, who wrote this sign? The tumults of my soul are stilled, My withered heart with rapture filled!	Ha! what delight does in a moment fill My senses at this sight! I feel at once The renovated streams of life and pleasure Bubble thro' every vein. Was it a god Who wrote this sign? it stills my soul's wild warfare; Fills my lost heart with joy, while some strange impulse	Spirits, ye that hover near, Speak and answer, if ye hear! Ha! what rapture from the sight Fills my veins with wild delight! Sure some God the sign has traced. In these features, plain and true, Nature's secrets greet my view.
501-509	In the currents of life, in tempests of motion, Hither and thither, Over and under, Wend I and wander — Birth and the grave— A limitless ocean, Where the restless wave Undulates ever— Under and over , Their toiling strife, I mingle and hover, The spirit of life;	In the floods of life, in the tempests of action, Up and down I rave; Hither and thither in motion; Birth and the grave, An unbounded ocean A changing strife A kindling life At the rustling loom of Time I have trod, And fashion'd the living vesture of God. Thou active spirit, circling the wide world,	I wander and range Through existence's change, Above and below, Through the tide and the flow, I shoot and I sparkle, and never am still. Say, thou ever- roving spirit, What relation can I bear to thee? To some other form, in another station, Thou mayest bear

		How near allied I feel myself to thee! SPIRIT. Thou'rt like the spirit	relation: Not to me. Not to thee! To whom then? I , the image of my Maker, Not to thee!
3217-3227	Yes! lofty spirit, thou hast given me all , All that I asked of thee; and not in vain Thy fiery countenance hast turned on me! —Hast given me empire o'er majestic nature, Power to enjoy and feel. 'Twas not alone The stranger's short permitted privilege Of momentary wonder, that thou gavest; No; thou hast given me into her deep breast As into a friend's secret heart to look; Hast brought to me the tribes of living things; Thus teaching me to recognise and love My brothers in still grove, or air, or stream. And when in the wide wood the tempest raves, And shrieks, and rends the giant pines, uproots,	Oh, thou great Spirit, thou hast given to me All, all that I desired. Thou hast not turned Thy beaming countenance in vain upon me. Thou gav'st me glorious Nature for a kingdom, The faculty to feel and to enjoy her. Thou didst not merely grant a cold short glimpse, But laid her deepest mysteries open to me , As a friend's bosom. All created things Thou mak'st to pass before me ; and the beings Peopling the fragile leaf— the air— the waters— Are to my sight revealed; while, when the storm	Not translated

	<p>Disbranches, and, with maddening grasp uplifting, Flings them to earth, and from the hollow hill Dull moaning thunders echo their descent; Then dost thou lead me to the safe retreat Of some low cavern, there exhibiting</p>	<p>Howls crackling through the forest—tearing down The giant pines, crushing both trunk and branch, And makes the hills re-echo to their fall, Then to the sheltering cave thou ledest me, And there layest bare the deep and secret places Of my own heart. There I may gaze upon</p>	
1675-1682	<p>What can'st thou give, poor miserable devil. Thinkest thou that man's proud soul— his struggling thoughts And high desires— have ever been conceived By such as thou art? wretch, what canst thou give? But thou hast food which satisfieth not, And thou hast the red gold, that restlessly Like quicksilver glides from the grasping hand And Play; at which none ever yet hath won,</p>	<p>Thou miserable fiend? can man's high spirit, Full of immortal longings, be by such As thou art, comprehended? Thou profferest food Which mocks its eager appetite; yellow gold, That melts like quicksilver in the grasping hand; Games at which none e'er won; enchanting woman, To lean upon my breast, and while she leans there</p>	Not Translated

As can be seen there are very remarkable function words agreement occurring not by simple coincidence in some of the passages of the 1821 *Faustus* translation and the two translations by Anster and Gowe: 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.

Chapter Five

Conclusions, Limitations, and Further Research

This study set out to test the hypothesis, proposed in Burwick and McKusick's re-edition of Boosey's 1821 *Faustus* translation, that Coleridge was the author of that translation. In this final chapter, we will present the conclusions that came from various clustering analyses, review the research contributions of this study, identify possible limitations in this study, as well as discuss directions for future research.

4.1 Conclusions:

The methodology used for testing was based on two fundamental principles of authorship attribution: that each author has a characteristic style which differentiates his or her work from that of others, and that an unassigned text can be attributed to the author whose style is most similar to, where similarity is measured on one or more stylistic criteria. The stylistic criterion used in this study was the frequency of usage of 80 function words by Coleridge and several contemporary authors on the one hand, and by the Boosey *Faustus* translation on the other.

The methodology was centred on the concept of the falsifiable hypothesis. 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 amend the hypothesis accordingly.

The hypothesis of Coleridgean authorship of the *Faustus* translation was tested by determining whether or not the *Faustus* translation's usage of the 80 function words was closer to that of Coleridge's usage than that of any other of the candidate authors included in the study. This determination was based on the concept of relative proximity of objects in high-dimensional vector space, where the objects in the present case are the texts included in the study, and the dimensionality of the vector space was determined by the 80 function words used as the stylistic criterion. Specifically, given a set $A = \{a_1, a_2, \dots, a_m\}$ of m texts by an author A and an anonymous text T :

- A set of n variables to describe the style of A and of T is selected. This defines an n -

dimensional vector space.

- T and each of the texts in A are measured in terms of those n variables and the results are stored in a matrix M with $(m+1)$ rows and n columns, such that the value at M_{ij} (for $i = 1..(m+1)$, $j = 1..n$) is the measurement of text i with respect to variable j .
- The values in each matrix row vector M_i (for $i = 1..(m+1)$) are the coordinates of a point in the vector space, and that point represents text i in the space.
- The relative similarity of any text i to any other is the distance between them.

The distances between the point representing T and the points representing the texts known to be by author A can then be interpreted as measures of stylistic similarity: if T is relatively far from those of A then the hypothesis that A is the author of T is falsified relative to the stylistic criteria used, and confirmed, though not of course proven, if relatively close.

Proximity in the vector space was measured between the texts on the basis of the frequency of usage of 80 function words using a range of cluster analytic methods: hierarchical clustering, principal component analysis, multidimensional scaling, Isomap, and the self-organizing map. The usages of 80 function words were identified in the texts by taking the means of the vectors using centroid analysis. The means of the vectors were bar-plotted and used to compare the Boosey *Faustus* translation to the Coleridge plays, as well as the plays by Shelley, Wordsworth, Byron and the five other translations of the play by Germaine de Staël, George Soane, Daniel Boileau, John Anster and Lord Francis Leveson-Gower.

The results from the various analyses showed that the 1821 *Faustus* was close to two other contemporary translations of the play by Anster and Gower (and in particular to Anster), but not to Coleridge's plays. The 1821 *Faustus* translation was also not close to those of the other three translations or the plays of Shelley, Wordsworth, and Byron.

The researcher believes that the different clustering results presented in the study allow a number of conclusions and raise some interesting possibilities for the anonymous 1821 *Faustus* translation authorship question:

1. The historical and, to some degree, the literary-critical evidence suggest Coleridgean 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. This yields:

- i) The hypothesis for Coleridge's authorship of the Boosey translation is falsified in favour of an alternative author. This result does not support Burwick and McKusick's claims. To the contrary, it strongly supports the opinion that many literary scholars hold against Coleridge's authorship or his candidacy as the true translator of the 1821 *Faustus*, as discussed in the Literature Review Chapter.
 - ii) Given that the result of the current study has previously been claimed by some literary scholars, what has been gained? The most obvious gain, as the researcher suggests, is confirmation: the obtained evidence from this study supports that literary claim and gives it a fresh scientific, objective and testable, ground. The implication is that the mathematical element in authorship attribution provides what Susan Hockey (2000:66) considers "concrete evidence to support or refute hypotheses or interpretations which have in the past been based on human reading and somewhat serendipitous noting of interesting features".
2. In terms of usages of 'and' and 'with' the 1821 *Faustus* and two other contemporary translations of the play by Anster and Gower are the most similar plays, while the other five Coleridge plays and the plays as well as the translations by the others differ strikingly in terms of 'that', 'to', 'then', 'from', 'and', and 'of' (higher or lower usages). The study does not make the claim that Anster or Gower translated the play being sensitive to the degree of incommensurability between the original work of Goethe and the 1821 translation (Boosey edition), and the additional interpretation presented here provides an explanation for this. The present study has identified many instances of function words borrowing by the anonymous translator of the 1821 *Faustus* and Gower from Anster's translation as well as by Gower from the 1821 translation and has also shown instances of using Anster's exact words and short phrases by these two authors (i.e. the anonymous translator and Gower). The study has further shown that Gower borrowed less frequently than the anonymous translator. Specifically, therefore, the present discussion can be said to have shown that borrowed function words from one author to another have the direct effect on clustering these three translations of *Faustus* all together.

3. Related to (2), the present study is the first to spot function words borrowing in *Faust* translations by the 1821 anonymous translator and Gower from Anster's translation. It is exceedingly surprising that a question of such obvious importance has been almost unnoticed by numerous studies devoted to the translation of Goethe's *Faustus*.

4. Whilst no claim for the 1821 *Faustus* translation to be a collaboration has ever been made and there is no stylometric method yet available that with accuracy and reliability approaches such a question, the study gives little attention to the hypothesis that Anster may have translated the 1821 *Faustus* in collaboration with Coleridge or Coleridge may have helped him in translating it. To date scholars do not know, but it is possible that they were working collaboratively, which would explain the claims made by Burwick and McKusick and other literary scholars who agreed with them that "verbal echoes and phrases in the translation connected in one manner or another with Coleridge's works...and plausible echoes and at times strong associations with Coleridge's poetry and plays", as discussed in Chapter One. Though all the works that bear Coleridge's name are all canonical works (i.e. well attributed and sole-authored), Coleridge is known to have collaborated on poetry and verse drama with Southey (e.g. *The Fall of Robespierre*) and Wordsworth (e.g. 'Lyrical Ballads'). If this was indeed the case or what happened with the 1821 *Faustus* translation, then the hypothesis just presented is consistent with Burwick's claims. According to Burwick, there is a connection between Coleridge and Anster as suggested in at least four occasions: (i) "John Anster had translated some 1,600 lines, closely imitating Goethe's varied verse forms, mixing them with just such as a prose analysis as Coleridge had proposed" (2007:xx); (ii) "In the opening scene, 'Night', he used exactly the same selection of passages as Anster" (2007: xxi); (iii) "Through the summer months of 1821, Anster continued to visit Highgate during his trips from Dublin to London" (2007: xxxiv), (iv) "The Huntington library has a copy of Anster's collected poems of 1819, with several corrections in Coleridge's hand as though he were trying to guide the young poet" (2007:xxxiii); and (v) "Coleridge's translation of *Faust* bears evidence of similar revisions of Anster's translation; evidence, that is that Coleridge had an eye on Anster's translation" (2007:xxxiii). The study does not exclude the possibility of collaboration, but currently we do not have much evidence to support either of them.

5. The study shows that cluster analysis is successful in distinguishing between several authors in a large corpus of texts on the basis of authorship style. Cluster analysis

methods are therefore recommended as effective methods for authorship attribution problems. Cluster analysis methods are also recommended for genre classification and forensic linguistics problems.

6. The study used mathematical techniques and procedures commonly related to the natural sciences such as Biology, Physics, Earth science, Neurology, etc. and fused them within the *Faust*-Coleridge authorship question, thereby brought into contact the two different fields, that is, mathematics and literature. Thus, the present study invites researchers to consider the following question from a variety of angles: what multivariate methods in general and cluster analysis in particular can offer to other areas of Humanities and Social Sciences research (e.g. Criminology, Political Sciences, Law, History, etc.) where data can be represented as a matrix, with the rows representing the objects to be clustered and the columns representing the variables that describe those objects?

4.2 Limitations:

It is important not to over-interpret the result advanced by the current study for the following reason. The study 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 hypothesis 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 hypothesis is abandoned.

4.3 Further research:

This study has thrown up two main questions related to the *Faust*-Coleridge authorship in need of further investigation. Based on the conclusions advanced in (2) and the caveat addressed in section (4.2) above, further work needs to be done in analysing other works by Anster and Gower using the same analytical methodology and different stylistic features such as word or character n-grams, which Grieve (2007) also rates well and perhaps others (e.g. Stamatatos, 2013; Luyckx, 2010; Koppel, et al., 2009; Houvardas and Stamatatos, 2006, 2008; Peng et al., 2004; Clement and Sharp, 2003; Keselj et al., 2003;

Ledger and Merriam, 1994) in order to establish whether Anster or Gower was the translator of an anonymous English 1821 translation of Goethe's German verse drama *Faustus* (Boosey edition).

Based on (4) above, when reliable stylometric analytical methods become available, further work also needs to be done on the question of collaboration between Coleridge and Anster to see whether Coleridge and Anster actually worked on the translation of the 1821 *Faustus* of Boosey's text together: This possibility should not be ruled out.

The researcher, based on the use of clustering analytical methods, remains convinced that scholars can not always assume that an individual who is attributed to a literary work was in fact the author.

Appendices
































The ASCII text files used in this study:

Appendix 1: 363 texts by Coleridge:

about nightingale1798	ass1794	comicauthor1825	duty1826	fears1798	friendchild1796	iambics1801
accident1826	atombless1809	concert room	easterholliday1787	fire1798	frost1798	ideas1830
adaydream1802	author1825	constancy1825	edward1825	forbearance1832	fulke1810	imitation1791
adventof love1824	authorpoems1795	CuriousND	effusion1792	foster1797	garden1828	immortalityofsoulND
advilum1805	autumnal1788	death1790	egerton1816	frag3ND	genevieve1790	improvisatore1827
ahymn1814	balahill1794	deathofchatterton1790	eminent1794	frag4ND	gentle1793	infant1794
akenside1794	ballad	deathofstarling1794	eolianharp1795	frag5ND	georgiana1799	inscription of seat
alamentia1805	baptisma1833	dejection1802	epigram59ND	frag6ND	german1799	inscription1802
alaysermon1817	baronND	delinquent1824	epigram64ND	frag7ND	germanpoet1802	insidethecoach1791
albumND	biographia1817	departing1796	epigram68ND	frag8ND	godmother1791	invitation1797
alice1828	birth1796	departure1828	epigram73ND	frag9ND	graves	invocation1790
alternative1825	Birthplace1799	deprofundis1806	epigram1806	frag10ND	grenville	irishlottery1793
am1798	blankverse1794	desire1830	epilogueND	frag12ND	happiness1791	israelament1817
amathematicalpro1791	Captin1804	destiny of nations	epitaphonhimself1803	frag13ND	happy husband	iyetremain1793
amelia1792	character1825	destructionofbastil1789	epitaphstop1833	frag14ND	hartz1799	julia1789
anallegoric1833	characters1794	devilthought1799	erskine	frag15ND	heracitus1822	keepsake
anangel1801	charityinthought1830	devonshireroad1791	eveningstar1790	frag18ND	hexameters1798	kepler1799
anexile1805	child question	disappointment1792	exchange1804	frag21	homeless1826	kettle1790
anna1790	childish	discovery 1794	experiment1801	frag191792	homesick	kiss1803
anthem1789	childprayer1806	domestic1794	Faded1794	frag201810	honor1791	kk1797
apologia1800	choleraND	dominici1826	faithandhope1815	france	hornet1796	knight1817
argument1811	christ1797	drinkingvsthinking1801	fancy1817	france1798	hourglass1811	knowledge1832
aristophanes1816	climbingBrockleyND	dungeon frag1798	farewell1791	frenzy1794	humalitythemother1830	ladybeauty1830
asoliloquy1800	coeli1830	dungeon1796	farewelltolove1805	friend no more poetry1796	humanlife1815	lastwordsofberengarius1826
asra	cologne1828	dural1787	farg11ND	friend1818	hymn to earth1799	late

lesbia1794	meetagain1795	nonsense3ND	phantomofact1830	river otter	sonnet1805	toafriend1795	whatislife18
lesson	melancholy	nose1789	piccolomini1800	Robbers1794	Spaces1800	toalady1822	willsND
lettercary1818	metricalfeet1806	nosense1ND	picture1802	robespierre1794	spring1794	toaladywithfalconer1814	winterND
lettercottle1814	missAT1828	notacriticND	Pity1795	rose1793	Stanhope1795	tomary1827	wish1792
lettertohenryND	missbarbour1829	notahome1830	pixes17933	rossettiND	stanzasND	tomycandle1802	woman at t
lettertojames1825	missbrunton1794	observing blossom1796	plaintive1814	ruinedhouse1797	stranger1800	tonature1820	workwitho
lettertopoole1801	moderncriticsND	ode1792	poetaster1796	sang a song1797	strophe1815	torevolution1795	xmascarol1
lettertosotheby1828	moon1788	ofhumane1810	presencelove1807	sappho1800	sunrise1802	tothenightingale1795	young frien
lettertothewall1796	orient1798	oldman1798	presentyear1833	school1791	sunset1805	towordsworth1807	young lady
lettertotulk1818	motto1808	on an infant1799	Priesty1794	sea shore	superstiti1794	tranquility1801	youngartist
lewti1798	muse1789	onaladyweeping1790	progress of vice	second birth1801	susan1829	translation1799	youngman
life1789	music1791	ondonne1818	prologue1794	secrecyND	tears1820	translation1826	youth1823
limbo1811	mutualpassionND	onpitt1806	prosestyle1818	sentimentalND	the day dream	translationofwrangham1794	zapolya1811
love1799	napoleon2ND	onreceivinganaccount1791	psyche1808	separation1805	the Wanderings of Cain1798	trochaics1808	
loveandfriendship1830	napoleonND	onseeing1791	quae1789	sheridan	thekingsarms1794	two founts1826	
loveburial1828	nativity1827	oscillators1798	rain1802	shurtonbars1795	thereproof1823	Unfortunate woman1797	
lovercomplaint1792	ne plus1826	osorio1797	rash1814	Siddons1794	thislime1797	unworthywisdomND	
luther1826	netherland1828	ossian1793	raven	sigh1794	thomashillND	virgin1811	
lyram1794	nightingale1798	Outcast1794	reason1830	Silver1795	thomaspooler1796	visionaryhope1810	
madman1809	nightmareND	pain1790	recantation1798	simileND	threesorts1835	visitofGod1799	
madmonk1800	nightscene1813	painsofsleep1803	recollections1807	Simplicity1797	time1812	walkbeforsupper1792	
Mahomet1799	nil pejus1787	pang1825	reflections1795	snowdrop1800	timepieceND	wallenstein1800	
manner1792	ninathomal1793	pantisocracy1794	religious musing	song1825	to a friend1794	water ballad 1799	
mannerofspencer1795	NonameND	pantisocracyinamerica1794	remorse1813	songbylovers1801	to a young lady1794	wellknownND	
marketclock1809	nonsense2ND	phantom1805	Rev G.Col1797	songs of Shepherds	to an infant	welsh1794	

Appendix 2: the 31 long texts by Coleridge:






















-  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

Appendix 3: the 332 short texts by Coleridge aggregated into 21 texts:

accident1826	authorpoems1795	deathofstarling1794	epigram73ND	frag10ND	hexameters1798	kettle1790
adaydream1802	alahill1794	departure1828	epigram1806	frag12ND	homeless1826	kiss1803
adventof love1824	ballad	deprofundis1806	epilogueND	frag13ND	homesick	kk1797
advilmum1805	baptisma1833	desire1830	epitaphonhimself1803	frag14ND	honor1791	knight1817
ahymn1814	baronND	destructionofbastil1789	epitaphstop1833	frag15ND	hornet1796	knowledge1832
akenside1794	biographia1817	deviltought1799	erskine	frag18ND	hourglass1811	ladybeauty1830
alament1805	birth1796	devonshireroad1791	eveningstar1790	frag21	humanitythemothel830	lastwordsofberengarius1826
alaysermon1817	Birthplace1799	disappointment1792	exchange1804	frag191792	humanlife1815	late
albumND	blankverse1794	discovery 1794	experiment1801	frag201810	hymn to earth1799	lesbia1794
alternative1825	Captin1804	domestic1794	Faded1794	france1798	iambics1801	lesson
amathematicalpro1791	character1825	dominic1826	faithandhope1815	frenzy1794	ideas1830	lettetocary1818
amelia1792	characters1794	drinkingvsthinking1801	fancy1817	friend no more poetry1796	imitation1791	lettetocottel1814
analogoric1833	charityinthought1830	dungeon frag1798	farewell1791	friendchild1796	immortalityofsoulND	lettetohenryND
anangel1801	child question	dungeon1796	farewelltolove1805	frost1798	infant1794	lettetojames1825
anexile1805	childish	dura1787	farg11ND	fulke1810	inscription of seat	lettetopool1801
anna1790	childprayer1806	duty1826	fire1798	garden1828	inscription1802	lettetosotheby1828
anthem1789	choleraND	easterholliday1787	forbearance1832	genevieve1790	insidethecoach1791	lettetothelwall1796
apologia1800	climbingBrockleyND	edward1825	foster1797	gentle1793	invitation1797	lettetotulk1818
argument1811	coeli1830	effusion1792	frag3ND	georgiana1799	invocation1790	lewti1798
aristophanes1816	cologne1828	egerton1816	frag4ND	german1799	irishlottery1793	life1789
asoliloquy1800	comicauthor1825	eminent1794	frag5ND	germanpoet1802	israellement1817	limbo1811
asra	concert room	eolianharp1795	frag6ND	godmother1791	iyetremain1793	love1799
ass1794	constancy1825	epigram59ND	frag7ND	happy husband	julia1789	loveandfriendship1830
atombless1809	CuriousND	epigram64ND	frag8ND	hartz1799	keepsake	loveburial1828
author1825	death1790	epigram68ND	frag9ND	heraclitus1822	kepler1799	lovercomplaint1792

luther1826	nightingale1798	painsofsleep1803	Robbers1794	spring1794	tomary1827	workwithouthope1825
lyram1794	nightmareND	pangl825	rose1793	Stanhope1795	tonycandle1802	xmscarol1799
madman1809	nightscene1813	pantisocracy1794	rossettiND	stanzasND	tonature1820	young friend
madmonk1800	nil pejus1787	pantisocracyinamerica1794	ruinedhouse1797	stranger1800	torevolution1795	young lady fever
Mahomet1799	ninathoma1793	phantom1805	sang a song1797	strophe1815	tothenightingale1795	youngartist1833
manner1792	NonameND	phantomorfact1830	sappho1800	sunrise1802	tranquility1801	youngman of fortune1796
mannerofspencer1795	nonsense2ND	Pity1795	school1791	sunset1805	translation1799	youth1823
marketclock1809	nonsense3ND	plaintive1814	sea shore	superstiti1794	translation1826	
meetagain1795	nose1789	poetaster1796	second birth1801	susan1829	translationofwranham1794	
melancholy	nosense1ND	presencelove1807	secrecyND	the day dream	trochaics1808	
missAT1828	notacriticND	presentyear1833	sentimentalND	the Wanderings of Cain1798	two founts1826	
missbarbour1829	notahome1830	Priestly1794	separation1805	thekingsarms1794	Unfortunate woman1797	
missbrunton1794	observing blossom1796	progress of vice	sheridan	thereproof1823	unworthywisdomND	
moderncriticsND	ode1792	prologue1794	shurtonbars1795	thislime1797	virgin1811	
moon1788	ofhumane1810	prosestyle1818	Siddons1794	thomashilND	visionaryhope1810	
moientil1798	on an infant1799	psychel1808	sight1794	thomaspoole1796	visitofGod1799	
motto1808	onaladyweeping1790	quae1789	Silver1795	threesorts1835	walkbeforsupper1792	
muse1789	ondonne1818	rain1802	simileND	time1812	water ballad 1799	
music1791	onpitt1806	rash1814	Simplicity1797	timepieceND	wellknownND	
mutualpassionND	onreceivinganaccount1791	raven	snowdrop1800	to a friend1794	welsh1794	
napoleon2ND	onseeing1791	reason1830	song1825	to a young lady1794	whatislifel1805	
napoleonND	oscillators1798	recollections1807	songbylovers1801	to an infant	willsND	
nativity1827	ossian1793	reflections1795	songs of Shepherds	toafriend1795	winterND	
ne plus1826	Outcast1794	Rev G.Col1797	sonnet1805	toalady1822	wish1792	
netherland1828	pain1790	river otter	Spaces1800	toaladywithfalconer1814	woman at theatre1797	

Appendix 4: the aggregated 21 texts by Coleridge:

-  Adaptations
-  An Old Man's Diary 1871
-  Anthology 1795
-  Biographia 1817
-  Cambridge intelligencer
-  Early Recollections 1837
-  Epigrams and Jeux D'esprit
-  Fragments
-  Friendship offering, New Mirror, magnet
-  Juvenile poems
-  Literary Remains 1836
-  Literary Souvenir
-  Lyrical Ballads 1798
-  Metrical Feet
-  Miscellaneous (later day)
-  Morning Chronicle
-  Morning Post
-  Sibylline Leaves 1817
-  The Courier
-  Unfinished Letters
-  Watchman







Appendix 5: 10 texts by Byron:

A – H (10)

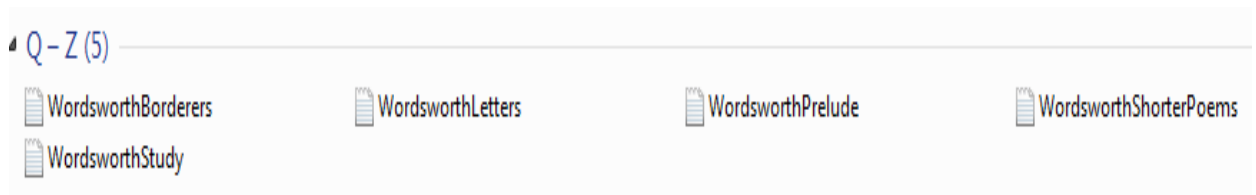
- | | | | |
|---|--|---|--|
|  Byron Cain |  Byron Deformed |  Byron Foscari |  Byron Harold |
|  Byron Heaven |  Byron Letters |  Byron Manfred |  Byron Sardanapulus |
|  Byron Shorter Poems |  Byron Werner | | |

Appendix 6: 6 texts by Shelley:

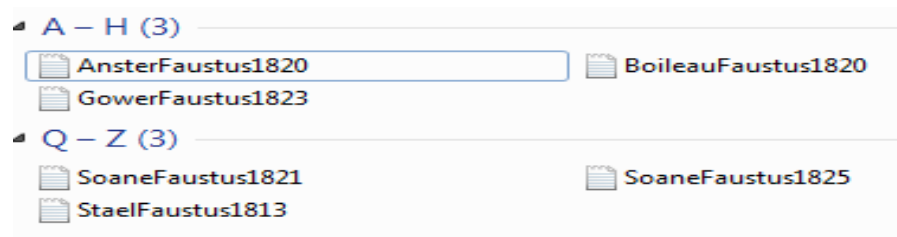
Q – Z (6)

- | | | | |
|--|---|--|---|
|  Shelley Adonais |  Shelley Cenci |  Shelley Defence |  Shelley Faust |
|  Shelley Prometheus |  Shelley Shorter Poems | | |

Appendix 7: 5 texts by Wordsworth:



Appendix 8: 5 texts by other *Faust* translators:

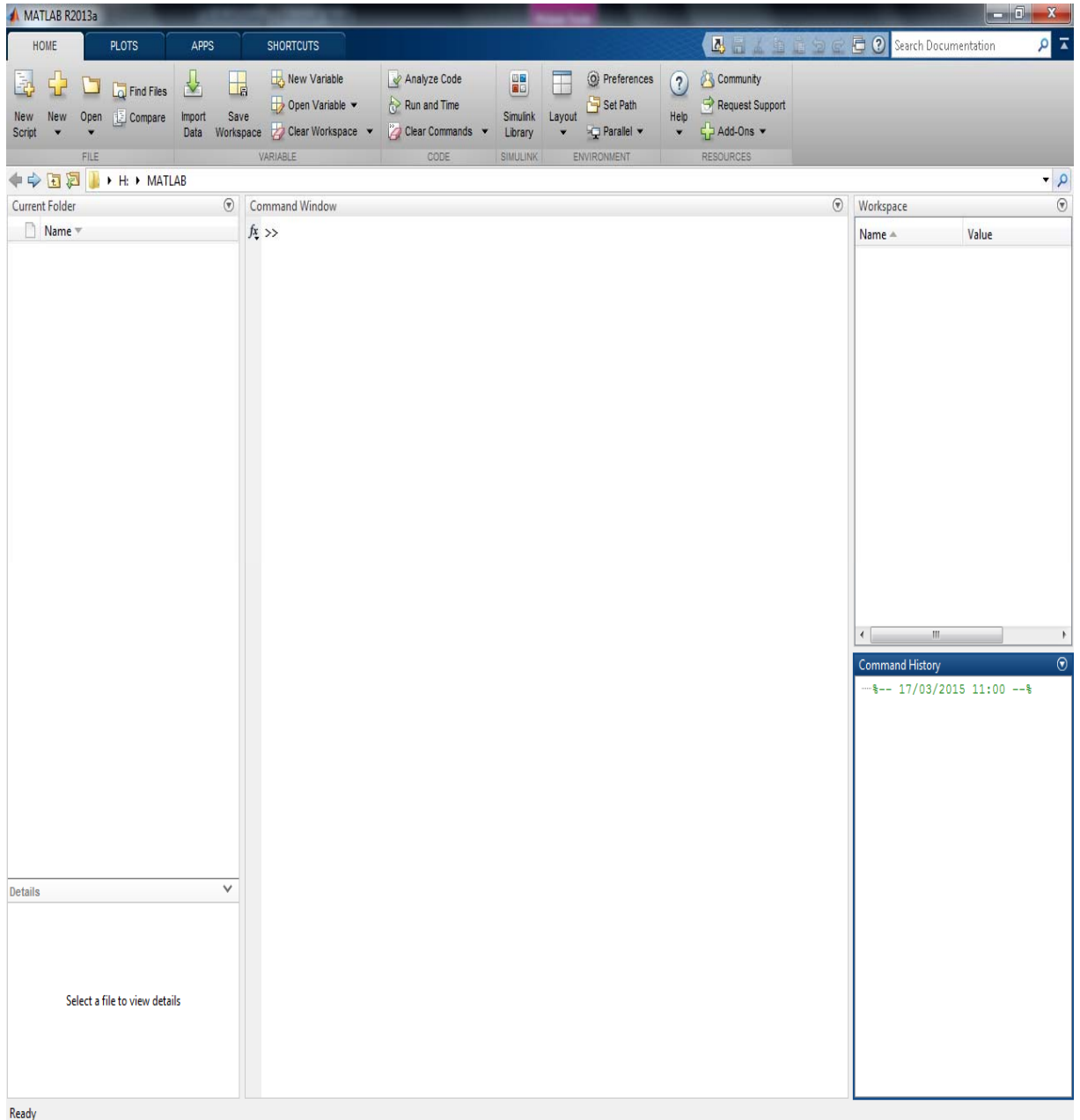


The programs used in the study:

Generation, adjusting, and analysis of the four data matrices (M80Norm, M180Norm, M280Nor, and M380Norm), on which this study is based, required different types of computational and data preparation programmes. Commercial and third-party tools such as MATALAB and CLUSTANGRAPHICS3 were used for the cluster analyses (hierarchical clustering, PCA, MDS, Isomap, and SOM) and for different types of graphics that present in the study. These are described in what follows:

Appendix 9: Matalab version R2013a:

MATLAB is a high-level numerical programming developed by MATHWORKS used for a very wide range of applications: data analysis, variable and matrix manipulations, plotting of data, etc. The user interface looks like this:



There are many functions in MATABL to select and describe, but all depend on the following basic steps for using it:

The desktop includes these panels:

- Current Folder: it is located on the left hand side where one can access his/her files.
- Command Window: it is the space located in the middle where you can enter commands at the command line, indicated by the prompt >>).
- Workspace: it is located on the right hand side where one can explore data that he/she generated or imported from files.

- Command History: it is located in the lower right angel where one can view or rerun commands that he/she entered at the command line.

In ways that are relevant to the present discussion, all cluster analysis results were generated using routines written in Matlab. Specifically:

1 PCA:

```
function [newmatrix, eigvect, eigval, variances] = pca (data, rotation, nroffactors)
```

```
% data is an  $m \times n$  matrix in which the  $m$  rows are observations and the  $n$ 
% columns are the variables.
```

```
% Rotation:
% 0 = no rotation
% 1 = varimax
```

```
% Determine size of matrix
[nrofrows nrofcols] = size (data);
```

```
% Calculate the column means for mean-centering
for j = 1:nrofcols
    sum = 0;
    for i = 1:nrofrows
        sum = sum + data(i,j);
    end
    colmeans(j) = sum / nrofrows;
end
```

```
% Mean-center the columns of the data matrix;
for j = 1:nrofcols
    for i = 1:nrofrows
        data(i,j) = data(i,j) - colmeans(j);
    end
end
```

```
% Get the covariance of the mean-centered data matrix
covdata = cov(data);
```

```
% Get the eigenvectors and eigenvalues
[evect eval] = eig(covdata);
```

```
% Abstract the variances from the eigenvalue matrix
variancevector = diag(eval);
```

```
% Matlab outputs the eigenvectors and eigenvalues in ascending rather than
% descending order of importance. Reverse-sort both the variance vector and
% the eigenvector matrix.
[junk, reverseindices] = sort(-1*variancevector);
```

```

variancevector = variancevector (reverseindices);
evect = evect(:,reverseindices);
eval = eval(:,reverseindices);

% Rotation
if rotation == 1
    evect = rotatefactors (evect, 'Method', 'varimax');
end

% Project the data into the new basis
for i = 1:nroffactors
    evectreduced(:,i) = evect(:,i);
end
proj = evectreduced' * data';

% Output
newmatrix = proj';
eigvect = evect;
eigval = eval;
variances = variancevector;

```

2. MDS:

[Y, stress] = mdscale [D,p], where

- *Y* is the output matrix
- *stress* is the stress measure associated with a particular result
- *D* is an $n \times n$ proximity matrix
- *p* is the dimensionality of the output space
- *mdscale* is the Matlab MDS routine

3 Isomap:

This is a Matlab script written by Tenenbaum & Langford and available for download at <http://isomap.stanford.edu/>.

- [Y, R, E] = Isomap(D, 'k', n), where:
 - *Y* is the output matrix
 - *R* is a vector of residual variances for embeddings in *Y*
 - *E* is the neighbourhood graph
 - *D* is a proximity matrix
 - '*k*' is the neighbourhood function
 - *n* is the neighbourhood size
 - *Isomap* is the Tenenbaum & Langford script

4 SOM:

- function varargout = som(varargin)
- % SOM Application M-file for som.fig
- % FIG = SOM launch som GUI.


```

• % SOM('callback_name', ...) invoke the named callback.
•
• if nargin == 0 % LAUNCH GUI
•
•     fig = openfig(mfilename,'reuse');
•
•     % Use system color scheme for figure:
•     set(fig,'Color',get(0,'defaultUicontrolBackgroundColor'));
•
•     % Generate a structure of handles to pass to callbacks, and store it.
•     handles = guihandles(fig);
•     guidata(fig, handles);
•
•     if nargin > 0
•         varargout{1} = fig;
•     end
•
• elseif ischar(varargin{1}) % INVOKE NAMED SUBFUNCTION OR
CALLBACK
•
•     try
•         [varargout{1:nargout}] = feval(varargin{:}); % FEVAL switchyard
•     catch
•         disp(lasterr);
•     end
•
• end
•
• % Standard SOM
• %
• % 1. GUI-controlled
•
• % 2. Data: reads from user-specified data file.
• %     Format:
• %     Line 1: nr of vectors
• %     Line 2: vector length
• %     Line 3 onwards: on each successive line, a numerical vector of length &
stated in line 2
•
• % 3. Architecture
• % a) 2-D map grid
• % b) Grid size user-specified. Defaults are built in, but can be changed to & %
any size (ie square or rectangular) via the GUI
• % c) Neighborhood shape, initial and final neighborhood, and rate of & %
neighborhood decrease user-specified. Defaults are
• %     built in, but can be changed via the GUI. Neighborhood decrease is %
linear, ie, decrease by 1 after n steps. Button for
• %     nonlinear decrease is on the GUI, but not yet implemented.
• % d) Learning rate initial and final values together with shape of learning %
rate decrease function. At the moment the only

```

- % option is linear = decrement after n steps; nonlinear is on the GUI, but % not yet implemented. Defaults are built in,
- % but can be changed via the GUI
- % e) Connection initialization can be either random or linear (ie, using % eigenvalues); at the moment only random
- % is implemented. Default is built in, but can be changed via the GUI
-
- % 4. Training
- % a) Sequential; no provision for batch
- % b) Selection of inputs is random using Matlab random number generator
- % c) Unit activation is by inner product. The GUI provides a normalization facility for input vectors.
- % d) Selection of best matching unit (BMU) is by Euclidean distance; distances are stored at each iteration in a distance matrix, which is used to identify the BMU = the smallest entry in the distance matrix at a given iteration.
- % Inner product is on the GUI but is not yet implemented
- % e) Neighborhood shape is currently either Diamond or Square. Details of these are given in the relevant section of the program, ie, the training routine
- % f) Weight update for a given grid unit (i,j) within neighborhood is proportional to distance of (i,j) from the unit of greatest activation. The proportion is $\delta / (\text{distance from selected unit})$, which, when plotted, gives a nonlinear decreasing curve as distance increases; applied to all the selected unit's neighbors.
- % g) Default number of training iterations is built in, but can be changed via the GUI
-
- % 5. Output
- % Two types:
- % a) Map lattice. Here the selection is 3-D mesh or surface plot, which can be rotated using Matlab graphics
- % b) Connection vectors: U-matrix
- % In (a), a single vector can be selected for plotting, or all the vectors can be plotted simultaneously
-
- % START PROGRAM
-
- % USEFUL FUNCTIONS
-
- % Euclidean distance between vectors
- function f = euclideandistance (v1, v2)
- [nrofrows nrofcols] = size (v1); % v2 must be same dimensionality
- sumofsquares = 0;
- for i = 1:nrofcols
- sumofsquares = sumofsquares + ((v1(i) - v2(i)) * (v1(i) - v2(i)));
- end
- f = sqrt (sumofsquares);

-
-
-
- `% DATA INPUT`
-
- `function varargout = radiobutton1_Callback(h, eventdata, handles, varargin)`
- `% Load data`
- `fname = inputdlg ('Name of input file');`
- `fname = fname {1}; % inputdlg returns a cell array; need to pull out the string, which is fname {1}`
- `handles.vectorlist = readvectfile (fname, '%f'); % Readvectfile is an externally-defined m-file`
- `% Store main dimensions of data`
- `[nrofdatavectors datavectorlength] = size (handles.vectorlist)`
- `handles.nrofdatavectors = nrofdatavectors;`
- `handles.datavectorlength = datavectorlength;`
- `handles.labellist = readlabelfile ('datalabels.txt'); % Readlabelfile is an externally-defined m-file`
- `guidata (gcbo, handles);`
-
-
- `% INITIALIZATION`
-
-
- `function radiobutton49_Callback(hObject, eventdata, handles)`
- `% Initialize constants`
- `% Output grid dimensions`
- `handles.mapnrofrows = 12;`
- `handles.mapnrofcols = 12;`
- `% Connection initialization`
- `handles.connectioninitialization = 'R'; %Random`
- `handles.connmin = 0.01; % Minimum random initialization value`
- `handles.connmax = 0.2; % Maximum`
- `handles.connectionnormalization = 'Y';`
- `% Neighborhood`
- `handles.initialneighborhood = 10;`
- `handles.finalneighborhood = 0;`
- `handles.neighborhooddecrementinterval = 50;`
- `handles.neighborhoodshape = 'S'; %Square`
- `handles.neighborhooddecrementmodel = 'L'; %Linear`
- `% Learning rate`
- `handles.initiallearningrate = 0.5;`
- `handles.finallearningrate = 0.011;`
- `handles.learningratedecrementstep = 0.01;`
- `handles.learningratedecrementinterval = 20;`
- `handles.learningratedecrementalgorithm = 'L'; %Linear`
- `% Best matching unit algorithm`
- `handles.bmualgorithm = 'E'; %Euclidean distance`
- `% Training iterations`
- `handles.nroftrainingiterations = 700;`
- `% Training display: show a 2-D map for each successive vector. Slows training down A LOT`
- `handles.displaygridduringtraining = 'N';`

- % Plots
- handles.superimposeplots = 'N';
- handles.displaylabels = 'N';
- handles.plotformat = 'S';
- handles.multipleplotrows = 3;
- handles.multipleplotcols = 2;
- handles.multipleplotnr = 1;
- guidata (gcbo,handles)
-
- % Display default values on GUI
- set (handles.edit1, 'string', int2str (handles.mapnrofrows));
- set (handles.edit2, 'string', int2str (handles.mapnrofcols));
- set (handles.edit3, 'string', int2str (handles.initialneighborhood));
- set (handles.edit4, 'string', int2str (handles.finalneighborhood));
- set (handles.edit5, 'string', int2str (handles.neighborhooddecrementinterval));
- set (handles.edit22, 'string', handles.neighborhoodshape);
- set (handles.edit26, 'string', handles.neighborhooddecrementmodel);
- set (handles.edit6, 'string', num2str (handles.initiallearningrate));
- set (handles.edit7, 'string', num2str (handles.finallearningrate));
- set (handles.edit8, 'string', num2str (handles.learningratedecrementstep));
- set (handles.edit9, 'string', int2str (handles.learningratedecrementinterval));
- set (handles.edit23, 'string', handles.learningratedecrementalgorithm);
- set (handles.edit25, 'string', handles.bmualgorithm);
- set (handles.edit19, 'string', int2str (handles.nroftrainingiterations));
- set (handles.edit20, 'string', handles.displaygridduringtraining);
- set (handles.edit24, 'string', handles.connectioninitialization);
- set (handles.edit27, 'string', handles.connectionnormalization);
- set (handles.edit28, 'string', handles.displaylabels);
- set (handles.edit29, 'string', num2str (handles.connmin));
- set (handles.edit30, 'string', num2str (handles.connmax));
- set (handles.edit35, 'string', handles.plotformat);
- guidata (gcbo, handles);
- set (handles.radiobutton17, 'fontweight', 'light', 'enable', 'off');
-
-
- % GUI PARAMETER SETTING
-
- function varargout = radiobutton10_Callback(h, eventdata, handles, varargin)
- % Square neighborhood shape
- handles.neighborhoodshape = 'S';
- set (handles.edit22, 'string', handles.neighborhoodshape);
- guidata (gcbo, handles);
-
- % -----
- function varargout = radiobutton11_Callback(h, eventdata, handles, varargin)
- % Diamond neighborhood shape
- handles.neighborhoodshape = 'D';
- set (handles.edit22, 'string', handles.neighborhoodshape);
- guidata (gcbo, handles);

```

•
• % -----
• function radiobutton53_Callback(hObject, eventdata, handles)
• % Spherical neighborhood shape
• handles.neighborhoodshape = 'P';
• set(handles.edit22, 'string', handles.neighborhoodshape);
• guidata(gcbo, handles);
•
• % -----
• function varargout = radiobutton12_Callback(h, eventdata, handles, varargin)
• % Display grid during training
• set(handles.edit20, 'string', 'Y');
•
• % -----
• function varargout = radiobutton13_Callback(h, eventdata, handles, varargin)
• % Display grid during training
• set(handles.edit20, 'string', 'N');
•
• % -----
• function varargout = radiobutton24_Callback(h, eventdata, handles, varargin)
• % Linear learning rate decrease algorithm
• handles.learningratedecrementalalgorithm = 'L';
• set(handles.edit23, 'string', handles.learningratedecrementalalgorithm);
• guidata(gcbo, handles);
•
• % -----
• function varargout = radiobutton26_Callback(h, eventdata, handles, varargin)
• % Nonlinear learning rate decrease algorithm
• disp('Nonlinear learning rate decrease algorithm not yet implemented')
•
• function varargout = radiobutton27_Callback(h, eventdata, handles, varargin)
• % Random initialization of connections
• handles.connectioninitialization = 'R';
• set(handles.edit24, 'string', handles.connectioninitialization);
• guidata(gcbo, handles);
•
• % -----
• function varargout = radiobutton28_Callback(h, eventdata, handles, varargin)
• % Linear initialization of connections
• disp('Linear initialization of connections not yet implemented')
•
• % -----
• function varargout = radiobutton29_Callback(h, eventdata, handles, varargin)
• % Superimpose individual vector plots
• handles.superimposeplots = 'Y';
• guidata(gcbo, handles);
•
• % -----
• function varargout = radiobutton36_Callback(h, eventdata, handles, varargin)

```

- % Set best matching unit algorithm to inner product
- handles.bmualgorithm = 'I';
- set(handles.edit25, 'string', handles.bmualgorithm);
- guidata(gcbo, handles);
-
- % -----
- function varargout = radiobutton37_Callback(h, eventdata, handles, varargin)
- % Set best matching unit algorithm to euclidean product
- handles.bmualgorithm = 'E';
- set(handles.edit25, 'string', handles.bmualgorithm);
- guidata(gcbo, handles);
-
- % -----
- function varargout = radiobutton38_Callback(h, eventdata, handles, varargin)
- % Set neighborhood decrement moden to linear
- handles.neighborhooddecrementmodel = 'L';
- set(handles.edit26, 'string', handles.neighborhooddecrementmodel);
- guidata(gcbo, handles);
-
- % -----
- function varargout = radiobutton39_Callback(h, eventdata, handles, varargin)
- % Set neighborhood decrement moden to nonlinear
- handles.neighborhooddecrementmodel = 'N';
- set(handles.edit26, 'string', handles.neighborhooddecrementmodel);
- guidata(gcbo, handles);
-
- % -----
- function varargout = radiobutton41_Callback(h, eventdata, handles, varargin)
- handles.connectionnormalization = 'Y';
- set(handles.edit27, 'string', handles.connectionnormalization);
- guidata(gcbo, handles);
-
- % -----
- function varargout = radiobutton42_Callback(h, eventdata, handles, varargin)
- handles.displaylabels = 'Y';
- set(handles.edit28, 'string', handles.displaylabels);
- guidata(gcbo, handles);
-
- % -----
- function varargout = radiobutton43_Callback(h, eventdata, handles, varargin)
- handles.displaylabels = 'N';
- set(handles.edit28, 'string', handles.displaylabels);
- guidata(gcbo, handles);
-
- % -----
- function radiobutton46_Callback(hObject, eventdata, handles)
- % Single plot
- handles.plotformat = 'S';
- set(handles.edit35, 'string', handles.plotformat);

- guidata (gcbo, handles);
-
- % -----
- function radiobutton47_Callback(hObject, eventdata, handles)
- % Multiple plot
- handles.plotformat = 'M';
- set (handles.edit35, 'string', handles.plotformat);
- handles.nrofmultipleplots = 1;
- guidata (gcbo, handles);
-
- % -----
- function varargout = edit1_Callback(h, eventdata, handles, varargin)
- % Nrofrows
- str = get (handles.edit1, 'string');
- handles.nrofrows = str2num (str);
- guidata (gcbo, handles);
-
- % -----
- function varargout = edit2_Callback(h, eventdata, handles, varargin)
- % Nrofcols
- str = get (handles.edit2, 'string');
- handles.nrofrows = str2num (str);
- guidata (gcbo, handles);
-
- % -----
- function varargout = edit3_Callback(h, eventdata, handles, varargin)
- % Initial neighborhood
- str = get (handles.edit3, 'string');
- handles.initialneighborhood = str2num (str);
- guidata (gcbo, handles);
-
- % -----
- function varargout = edit4_Callback(h, eventdata, handles, varargin)
- % Final neighborhood
- str = get (handles.edit4, 'string');
- handles.finalneighborhood = str2num (str);
- guidata (gcbo, handles);
-
- % -----
- function varargout = edit5_Callback(h, eventdata, handles, varargin)
- % Neighborhhod decrement interval
- str = get (handles.edit5, 'string');
- handles.neighborhooddecrementinterval = str2num (str);
- guidata (gcbo, handles);
-
- % -----
- function varargout = edit6_Callback(h, eventdata, handles, varargin)
- % Initial learning rate
- str = get (handles.edit6, 'string');

- handles.initiallearningrate = str2num (str);
- guidata (gcbo, handles);
-
- % -----
- function varargout = edit7_Callback(h, eventdata, handles, varargin)
- % Final learning rate
- str = get (handles.edit7, 'string');
- handles.finalllearningrate = str2num (str);
- guidata (gcbo, handles);
-
- % -----
- function varargout = edit8_Callback(h, eventdata, handles, varargin)
- % Learning rate decrement step
- str = get (handles.edit8, 'string');
- handles.learningratedecrementstep = str2num (str);
- guidata (gcbo, handles);
-
- % -----
- function varargout = edit9_Callback(h, eventdata, handles, varargin)
- % Learning rate decrement interval
- str = get (handles.edit9, 'string');
- handles.learninratedecrementinterval = str2num (str);
- guidata (gcbo, handles);
-
- % -----
- function varargout = edit12_Callback(h, eventdata, handles, varargin)
- % Get single input vector to generate a map for it after training
- str = get (handles.edit12, 'string');
- handles.selectedinputvectorindex = str2num (str);
- guidata (gcbo, handles);
-
- % -----
- function varargout = edit19_Callback(h, eventdata, handles, varargin)
- % Number of training iterations
- str = get (handles.edit19, 'string');
- handles.nroftrainingiterations = str2num (str);
- guidata (gcbo, handles);
-
- % -----
- function varargout = edit20_Callback(h, eventdata, handles, varargin)
- str = get (handles.edit20, 'string');
- handles.displaygridduringtraining = str2num (str);
- guidata (gcbo, handles);
-
-
-
- % TRAIN SOM
-
- % -----
- function varargout = radiobutton6_Callback(h, eventdata, handles, varargin)

- % Training
-
- % Initialize data parameters. Any user changes to GUI-displayed defaults are captured here
- handles.nrofrows = str2num (get (handles.edit1, 'string'));
- handles.nrofcols = str2num (get (handles.edit2, 'string'));
- handles.nrofunits = handles.nrofrows * handles.nrofcols;
- handles.initialneighborhood = str2num (get (handles.edit3, 'string'));
- handles.finalneighborhood = str2num (get (handles.edit4, 'string'));
- handles.neighborhooddecrementinterval = str2num (get (handles.edit5, 'string'));
- handles.neighborhoodshape = get (handles.edit22, 'string');
- handles.neighborhooddecrementmodel = get (handles.edit26, 'string');
- handles.learningratedecrementalgorithm = get (handles.edit23, 'string');
- handles.initiallearningrate = str2num (get (handles.edit6, 'string'));
- handles.finallearningrate = str2num (get (handles.edit7, 'string'));
- handles.learningratedecrementstep = str2num (get (handles.edit8, 'string'));
- handles.learningratedecrementinterval = str2num (get (handles.edit9, 'string'));
- handles.bmualgorithm = get (handles.edit25, 'string');
- handles.nroftrainingiterations = str2num (get (handles.edit19, 'string'));
- handles.displaygridduringtraining = get (handles.edit20, 'string');
- handles.connectioninitialization = get (handles.edit24, 'string');
- handles.connectionnormalization = get (handles.edit27, 'string');
- handles.displaylabels = get (handles.edit28, 'string');
- handles.connmin = str2num (get (handles.edit29, 'string'));
- handles.connmax = str2num (get (handles.edit30, 'string'));
- handles.plotformat = get (handles.edit35, 'string');
- handles.multipleplotrows = handles.multipleplotrows;
- handles.multipleplotcols = handles.multipleplotcols;
- guidata (gcbo, handles);
-
- % Random initialization of connections. These are stored in a 2-D matrix: there are as many rows as units in the map, and
- % row [i] is the connection vector for unit [i]
- if handles.connectioninitialization == 'R'
- % See Matlabd Help 'rand' for this
- handles.connections = handles.connmin + (handles.connmax - handles.connmin) * rand ([handles.nrofunits handles.datavectorlength])
- end
-
- % Linear initialization of connections. These are stored in a 2-D matrix: there are as many rows as units in the map, and
- % row [i] is the connection vector for unit [i]
- if handles.connectioninitialization == 'L'
- disp ('Linear sonnection initialization not yet implemented');
- end;
-
- % If specified, normalize the connection vectors to interval 0..1
- if handles.connectionnormalization == 'Y'
- % Minimum value of original connection vector

- origvectorminvalue = handles.connmin;
- % Minimum value of normalized connection vector
- normvectorminvalue = 0;
- % Maximum value of normalized connection vector
- normvectormaxvalue = 1;
- % For each data vector in turn
- for i = 1:handles.nrofunits
- % Get the next vector
- currentconnvector = handles.connections (i,:);
- % Initialize the maximum value in the current data vector
- origvectormaxvalue = 0;
- % Go through the current data vector to get the largest value
- for j = 1:handles.datavectorlength
- if handles.connections (i,j) > origvectormaxvalue
- origvectormaxvalue = handles.connections (i,j);
- end
- end
- % Normalize; formula given by code
- for j = 1:handles.datavectorlength
- handles.connections (i,j) = ((currentconnvector (j)- origvectorminvalue) *
 (normvectormaxvalue - normvectorminvalue) /...
 (origvectormaxvalue - origvectorminvalue)) +
 normvectorminvalue;
- end
- end
- end
-
- % Initialize map grid
- handles.grid = zeros (handles.nrofrows, handles.nrofcols);
- % Initialize distance matrix
- handles.distancematrix = zeros (handles.nrofrows, handles.nrofcols);
- % Initialize local variables
- currentneighborhood = handles.initialneighborhood;
- currentlearningrate = handles.initiallearningrate;
-
- guidata (gcbo, handles);
-
- % Start training
- for i = 1:handles.nroftrainingiterations
-
- % Respond to change in map display button
- handles.displaygridduringtraining = get (handles.edit20, 'string');
-
- % GUI output
- set (handles.edit13, 'string', num2str (i));
- drawnow;
-
- % Randomly select an input vector. Essentially, keep generating until a number in
 the <= the number of training

- % vectors is generated
- indexfound = 0;
- while indexfound == 0
- x = round (rand * 100);
- if x == 0
- x = 1;
- end
- if x <= handles.nrofdatavectors
- vectorindex = x;
- indexfound = 1;
- end
- end
- % Select the vector using the generated index
- currentdatavector = handles.vectorlist (vectorindex,:);
- % GUI output
- set (handles.edit14, 'string', num2str (vectorindex));
- drawnow;
-
- % Generate a distance matrix from the match between the current data vector and each connection vector in turn. This
- % can be done using either Euclidean distance or inner product.
- for j = 1:handles.nrofrows
- for k = 1:handles.nrofcols
- % For a given (j,k) unit, find the index to the associated weight vector in the weight matrix. This is done using offset, as below
- connectionindex = (handles.nrofrows * (j - 1)) + k;
- % Use the index to get the weight vector of (j,k)
- connectionvector = handles.connections (connectionindex,:);
-
- % If using Euclidean distance of data and connection vectors to generate distance matrix
- if handles.bmualgorithm == 'E'
- sumofsquares = 0;
- for m = 1:handles.datavectorlength
- sumofsquares = sumofsquares + ((currentdatavector (m) - connectionvector (m)) * (currentdatavector (m) - connectionvector (m)));
- end
- handles.distancematrix (j,k) = sqrt (sumofsquares);
- end
-
- % If using inner product of data and connection vectors to generate distance matrix
- if handles.bmualgorithm == 'I'
- innerproduct = 0;
- for m = 1:handles.datavectorlength
- innerproduct = innerproduct + (currentdatavector (m) * connectionvector (m));
- end
- handles.distancematrix (j,k) = innerproduct;
- end

- end
- end
-
- % Find the BMU
- % If Euclidean distance activation was used, we are looking for the smallest value in the distance matrix
- if handles.bmualgorithm == 'E'
- smallestactivation = 100000;
- for j = 1:handles.nrofrows
- for k = 1:handles.nrofcols
- if handles.distancematrix (j,k) < smallestactivation
- selectedcell = [j k];
- smallestactivation = handles.distancematrix (j,k);
- end
- end
- end
- end
-
- % If inner product activation was used, we are looking for the largest value in the distance matrix
- if handles.bmualgorithm == 'I'
- largestactivation = 0;
- for j = 1:handles.nrofrows
- for k = 1:handles.nrofcols
- if handles.distancematrix (j,k) > largestactivation
- selectedcell = [j k];
- largestactivation = handles.distancematrix (j,k);
- end
- end
- end
- end
-
- % GUI output
- set(handles.edit15, 'string', num2str(selectedcell));
- drawnow;
-
- % DIAMOND NEIGHBORHOOD
- if handles.neighborhoodshape == 'D'
- % The first step is to identify the units in the neighborhood of the selected unit, given the current neighborhood size.
- % Implementation:
- % 1. Identify the start and end rows that, given the current neighborhood, bound the selected unit (account is taken of
- % boundary cases, where the selected unit is placed relative to the top or bottom map boundary in such a way that the full
- % neighborhood is not possible.
- % 2. The number of left and right units around the selected unit depends on the row-distance from the selected unit. For
- % example, say the selected unit is at location (6,4), where 6 is the row, and the current neighborhood is 3. The start

- % row is thus 3, and the end row 9. Going from the top, and keeping in mind that the neighborhood is diamond shaped, the
- % only unit in the neighborhood is (3,4), ie, there are no units to the left and right of the one in the (4) column above
- % the selected unit. At (4,4), there is one unit to the left and one unit to the right of the (4)-column, at (5,4) there are
- % two to the left and two to the right, at (6,4) 3 to the left and 3 to the right. At (7,4), again keeping in mind the
- % diamond shape, the left and right units begin to decrease again: 2 1 0.
- %
- % For each unit thus identified, do the necessary connection vector update using the SOM algorithm, based on the distance of the unit
- % from the selected one (for details see below)
-
- startrow = selectedcell (1) - currentneighborhood;
- % Boundary condition
- if startrow < 1
- startrow = 1;
- end
-
- endrow = selectedcell (1) + currentneighborhood;
- % Boundary condition
- if endrow > handles.nrofrows
- endrow = handles.nrofrows;
- end;
-
- % For each row bounding the selected unit
- for j = startrow:endrow
- % Determine how many units left and right given the row (including boundary conditions, where full left or right context
- % might not be possible
-
- % Above and including the row of the selected cell, the number of left and right units increases as one works from
- % the top row downwards
- if j <= selectedcell (1)
- left = selectedcell (2) - (currentneighborhood - (selectedcell (1) - j));
- if left < 1
- left = 1;
- end
- right = selectedcell (2) + (currentneighborhood - (selectedcell (1) - j));
- if right > handles.nrofcols
- right = handles.nrofcols;
- end
- else
- % Below the selected cell row, however, the number of left and right units decreases as the row number increases
- left = selectedcell (2) - (currentneighborhood - (j - selectedcell (1)));
- if left < 1
- left = 1;

- end
- $\text{right} = \text{selectedcell}(2) + (\text{currentneighborhood} - (j - \text{selectedcell}(1)))$;
- if $\text{right} > \text{handles.nrofcols}$
- $\text{right} = \text{handles.nrofcols}$;
- end
- end
-
- % Given the current neighborhood row, and the number of left-right units, carry out the SOM algorithm on each unit on the row
- for $k = \text{left}:\text{right}$
- % Calculate the index for the connection vector associated with the current unit using offset of the current unit's (row, column)
- % location in the map grid.
- $\text{connectionindex} = (\text{handles.nrofcols} * (j - 1)) + k$;
- % Using that index, get the connection vector
- $\text{connectionvector} = \text{handles.connections}(\text{connectionindex},:)$;
- % Get the difference between corresponding input and connection vectors, and multiply that difference by the current learning rate
- for $m = 1:\text{handles.datavectorlength}$
- $\text{connectionvector}(m) = \text{connectionvector}(m) + (\text{currentlearningrate} * (\text{currentdatavector}(m) - \text{connectionvector}(m)))$;
- end
- % Put the updated vector back into the main connection vector list
- $\text{handles.connections}(\text{connectionindex},:) = \text{connectionvector}$;
- end
- end
- end
-
- % SQUARE NEIGHBORHOOD
- if $\text{handles.neighborhoodshape} == 'S'$
- % The first step is to identify the units in the neighborhood of the selected unit, given the current neighborhood size.
- % Implementation:
- % 1. Identify the start and end rows that, given the current neighborhood, bound the selected unit (account is taken of
- % boundary cases, where the selected unit is placed relative to the top or bottom map boundary in such a way that the full
- % neighborhood is not possible.
- % 2. The number of left and right units around the selected unit depends on the row-distance from the selected unit. For
- % example, say the selected unit is at location (6,4), where 6 is the row, and the current neighborhood is 3.
- % a) Rows: the start row is $6 - 3 = 3$ and the endrow is $6 + 3 = 9$
- % b) Cols: the start col is $4 - 3 = 1$ and the end col is $4 + 3 = 7$
- % For each unit thus identified, do the necessary connection vector update using the SOM algorithm, based on the distance of the unit
- % from the selected one (for details see below)
-
- $\text{startrow} = \text{selectedcell}(1) - \text{currentneighborhood}$;
- % Boundary condition

- if startrow < 1
- startrow = 1;
- end
-
- endrow = selectedcell (1) + currentneighborhood;
- % Boundary condition
- if endrow > handles.nrofrows
- endrow = handles.nrofrows;
- end;
-
- left = selectedcell (2) - currentneighborhood;
- % Boundary condition
- if left < 1
- left = 1;
- end
-
- right = selectedcell (2) + currentneighborhood;
- % Boundary condition
- if right > handles.nrofcols
- right = handles.nrofcols;
- end
-
- % Given the above row & column boundaries, carry out the SOM algorithm on each unit within the neighborhood
- for j = startrow:endrow
- for k = left:right
- % Calculate the index for the connection vector associated with the current unit using offset of the current unit's (row, column)
- % location in the map grid.
- connectionindex = (handles.nrofcols * (j - 1)) + k;
- % Using that index, get the connection vector
- connectionvector = handles.connections (connectionindex,:);
- % Get the difference between corresponding input and connection vectors, and multiply that difference by the current learning rate
- for m = 1:handles.datavectorlength
- connectionvector (m) = connectionvector (m) + (currentlearningrate * (currentdatavector (m) - connectionvector (m)));
- end
- % Put the updated vector back into the main connection vector list
- handles.connections (connectionindex,:) = connectionvector;
- end
- end
- end
-
- % SPHERICAL NEIGHBORHOOD
- if handles.neighborhoodshape == 'P'
- % The first step is to identify the units in the neighborhood of the selected unit, given the current neighborhood size.
- % Implementation:

- % 1. Identify the start and end rows that, given the current neighborhood, bound the selected unit (account is taken of
- % boundary cases, where the selected unit is placed relative to the top or bottom map boundary in such a way that the full
- % neighborhood is not possible.
- % 2. The number of left and right units around the selected unit depends on the row-distance from the selected unit. For
- % example, say the selected unit is at location (6,4), where 6 is the row, and the current neighborhood is 3.
- % a) Rows: the start row is $6-3 = 3$ and the endrow is $6 + 3 = 9$
- % b) Cols: the start col is $4 - 3 = 1$ and the end col is $4 + 3 = 7$
- % For each unit thus identified, do the necessary connection vector update using the SOM algorithm, based on the distance of the unit
- % from the selected one (for details see below)
-
- startrow = selectedcell (1) - currentneighborhood;
- % Boundary condition
- if startrow < 1
- startrow = handles.nrofrows - abs (startrow);
- end
-
- endrow = selectedcell (1) + currentneighborhood;
- % Boundary condition
- if endrow > handles.nrofrows
- endrow = endrow - handles.nrofrows;
- end;
-
- left = selectedcell (2) - currentneighborhood;
- % Boundary condition
- if left < 1
- left = handles.nrofcols - abs (left);
- end
-
- right = selectedcell (2) + currentneighborhood;
- % Boundary condition
- if right > handles.nrofcols
- right = right - handles.nrofcols;
- end
-
- % Given the above row & column boundaries, carry out the SOM algorithm on each unit within the neighborhood
- for j = startrow:endrow
- for k = left:right
- % Calculate the index for the connection vector associated with the current unit using offset of the current unit's (row, column)
- % location in the map grid.
- connectionindex = (handles.nrofcols * (j - 1)) + k;
- % Using that index, get the connection vector
- connectionvector = handles.connections (connectionindex,:);
- % Get the difference between corresponding input and connection vectors, and


```

multiply that difference by the current learning rate
•   for m = 1:handles.datavectorlength
•   connectionvector (m) = connectionvector (m) + (currentlearningrate *
(currentdatavector (m) - connectionvector (m)));
•   end
•   % Put the updated vector back into the main connection vector list
•   handles.connections (connectionindex,:) = connectionvector;
•   end
•   end
•   end
•
•   % For debugging or observation it may be useful to show the map grid at each
training step
•   if handles.displaygridduringtraining == 'Y'
•   % Generate a map grid for the current data vector
•   for j = 1:handles.nrofrows
•   for k = 1:handles.nrofcols
•   connectionindex = (handles.nrofcols * (j - 1)) + k;
•   connectionvector = handles.connections (connectionindex,:);
•   innerproduct = dot (currentdatavector, connectionvector);
•   handles.grid (j,k) = innerproduct;
•   end
•   end
•   % Define where the figure showing the grid will be
•   rect = [370 95 500 510];
•   % create the grid figure in that position
•   handles.mapgrid = figure ('position', rect);
•   % Do a surface plot of the current activation grid of the map
•   surf (handles.grid);
•   % Show the map in (row, column) format with row 1 at the top left
•   axis ij;
•   %Label the axes
•   axis ([1 handles.nrofrows 1 handles.nrofcols]);
•   % 2-D vertical view
•   view (0,90);
•   % Pause to give time to look at the grid
•   pause (5);
•   % Delete the grid
•   delete (handles.mapgrid);
•   end
•
•   % If it's time to decrease the neighborhood, do that
•   if (mod (i,handles.neighborhooddecrementinterval) == 0) &
(currentneighborhood > handles.finalneighborhood)
•   currentneighborhood = currentneighborhood - 1;
•   end;
•   set (handles.edit16, 'string', num2str (currentneighborhood));
•   drawnow;
•

```


- % Define the plot figure's location
- rect = [370 95 500 510];
- % Create the plot figure using that location
- handles.mapgrid = figure ('position', rect);
- end
-
- % -----
- function radiobutton44_Callback(hObject, eventdata, handles)
- % Global plot: BMU only
-
- % Initialize the BMU map to all zeros; this will then be updated as
- % vectors are presented
- handles.grid = zeros (handles.nrofrows, handles.nrofcols);
- %This matrix keeps track of how many vectors are assigned to a given map
- %cell. It is used to adjust where on the display the label is placed
- handles.cellselectionmatrix = zeros (handles.nrofrows, handles.nrofcols);
-
- % For each vector in turn
- for i = 1:handles.nrofdatavectors
- % Get the current vector
- smallest = 100000;
- currentvector = handles.vectorlist (i,:);
- % Generate a map for the current vector
- for j = 1:handles.nrofrows
- for k = 1:handles.nrofcols
- connectionindex = (handles.nrofcols * (j - 1)) + k;
- connectionvector = handles.connections (connectionindex,:);
- distance = euclideanDistance (currentvector, connectionvector);
- if distance < smallest
- row = j;
- col = k;
- smallest = distance;
- end
- end
- end
- end
- % Mark this cell as most-active
- handles.grid (row,col) = 1;
- %Note that a vector is assigned to this cell
- handles.cellselectionmatrix (row,col) = handles.cellselectionmatrix (row,col) + 1;
- mesh (handles.grid);
- hold on;
- if handles.cellselectionmatrix (row,col) == 1
- t = text (row,col, handles.labellist (i), 'VerticalAlignment', 'bottom', 'FontSize', 10);
- end
- if handles.cellselectionmatrix (row,col) == 2
- t = text (row,col, handles.labellist (i), 'VerticalAlignment', 'middle', 'FontSize', 10);
- end

- if handles.cellselectionmatrix (row,col) > 2
- t = text (row,col, handles.labellist (i), 'VerticalAlignment', 'top', 'FontSize', 10);
- end
- end
-
- % Show the map in (row, column) format with row 1 at the top left
- axis xy;
- % Label axes
- axis ([1 handles.nrofcols 1 handles.nrofrows]);
- axis off;
- view (90,90);
- if handles.plotformat == 'M'
- handles.multipleplotnr = handles.multipleplotnr + 1;
- end;
- guidata (gcbo, handles);
-
- % -----
- function varargout = radiobutton30_Callback(h, eventdata, handles, varargin)
- % Global plot: mesh
-
- %This matrix keeps track of how many vectors are assigned to a given map
- %cell. It is used to adjust where on the display the label is placed
- handles.cellselectionmatrix = zeros (handles.nrofrows, handles.nrofcols);
- for i = 1:handles.nrofdatavectors
- smallest = 100000;
- % Get the current vector
- currentvector = handles.vectorlist (i,:);
- % Generate a map for the current vector
- for j = 1:handles.nrofrows
- for k = 1:handles.nrofcols
- connectionindex = (handles.nrofcols * (j - 1)) + k;
- connectionvector = handles.connections (connectionindex,:);
- distance = euclideanDistance (currentvector, connectionvector);
- if distance < smallest
- row = j;
- col = k;
- smallest = distance;
- end
- handles.grid (i,j) = distance;
- end
- end
- %Note that a vector is assigned to this cell
- handles.cellselectionmatrix (row,col) = handles.cellselectionmatrix (row,col) + 1;
-
- if handles.plotformat == 'M'
- subplot (handles.multipleplotrows, handles.multipleplotcols, handles.multipleplotnr);
- end
- mesh (handles.grid);

- if handles.cellselectionmatrix (row,col) == 1
- t = text (row, col, handles.labellist (i), 'VerticalAlignment', 'bottom', 'FontSize', 10);
- end
- if handles.cellselectionmatrix (row,col) == 2
- t = text (row, col, handles.labellist (i), 'VerticalAlignment', 'middle', 'FontSize', 10);
- end
- if handles.cellselectionmatrix (row,col) > 2
- t = text (row, col, handles.labellist (i), 'VerticalAlignment', 'top', 'FontSize', 10);
- end
- hold on;
- end
- % Show the map in (row, column) format with row 1 at the top left
- axis ij;
- % Label axes
- axis ([1 handles.nrofrows 1 handles.nrofcols]);
- axis off;
- view (90, 90);
- if handles.plotformat == 'M'
- handles.multipleplotnr = handles.multipleplotnr + 1;
- end
- guidata (gcbo, handles);
-
- % -----
- function radiobutton45_Callback(hObject, eventdata, handles)
- % Global plot: surface
-
- % Initialize the label list
- labellistlength = 0;
- % For each vector in turn
- handles.cellselectionmatrix = zeros (handles.nrofrows, handles.nrofcols);
- for i = 1:handles.nrofdatavectors
- smallest = 100000;
- % Get the current vector
- currentvector = handles.vectorlist (i,:);
- % Generate a map for the current vector
- for j = 1:handles.nrofrows
- for k = 1:handles.nrofcols
- connectionindex = (handles.nrofcols * (j - 1)) + k;
- connectionvector = handles.connections (connectionindex,:);
- distance = euclideandistance (currentvector, connectionvector);
- if distance < smallest
- row = j;
- col = k;
- smallest = distance;
- end
- handles.grid (j,k) = distance;
- end

- end
-
- %Note that a vector is assigned to this cell
- handles.cellselectionmatrix (row,col) = handles.cellselectionmatrix (row,col) + 1;
-
- if handles.plotformat == 'M'
- subplot (handles.multipleplotrows, handles.multipleplotcols, handles.multipleplotnr);
- end
- surf (handles.grid);
- if handles.cellselectionmatrix (row,col) == 1
- t = text (row, col, handles.labellist (i), 'VerticalAlignment', 'bottom', 'FontSize', 10);
- end
- if handles.cellselectionmatrix (row,col) == 2
- t = text (row, col, handles.labellist (i), 'VerticalAlignment', 'middle', 'FontSize', 10);
- end
- if handles.cellselectionmatrix (row,col) > 2
- t = text (row, col, handles.labellist (i), 'VerticalAlignment', 'top', 'FontSize', 10);
- end
- hold on;
- end
-
- % Show the map in (row, column) format with row 1 at the top left
- axis ij;
- % Label axes
- axis ([1 handles.nrofrows 1 handles.nrofcols]);
- axis off;
- view (90, 90);
- % Interpolated
- shading interp;
- if handles.plotformat == 'M'
- handles.multipleplotnr = handles.multipleplotnr + 1;
- end;
- guidata (gcbo, handles);
-
- % -----
- function radiobutton50_Callback(hObject, eventdata, handles)
- % Global plot: contour
-
- % For each vector in turn
- for i = 1:handles.nrofdatavectors
- min = 100000;
- % Get the current vector
- smallest = 100000;
- currentvector = handles.vectorlist (i,:);
- handles.cellselectionmatrix = zeros (handles.nrofrows, handles.nrofcols);
- % Generate a map for the current vector

- for j = 1:handles.nrofrows
- for k = 1:handles.nrofcols
- connectionindex = (handles.nrofcols * (j - 1)) + k;
- connectionvector = handles.connections (connectionindex,:);
- distance = euclideanistance (currentvector, connectionvector);
- if distance < smallest
- row = j;
- col = k;
- smallest = distance;
- end
- handles.grid (j,k) = distance;
- end
- end
-
- %Note that a vector is assigned to this cell
- handles.cellselectionmatrix (row,col) = handles.cellselectionmatrix (row,col) + 1;
-
- if handles.plotformat == 'M'
- subplot (handles.multipleplotrows, handles.multipleplotcols, handles.multipleplotnr);
- end
- contour (handles.grid);
- if handles.cellselectionmatrix (row,col) == 1
- t = text (row, col, handles.labellist (i), 'VerticalAlignment', 'bottom', 'FontSize', 10);
- end
- if handles.cellselectionmatrix (row,col) == 2
- t = text (row, col, handles.labellist (i), 'VerticalAlignment', 'middle', 'FontSize', 10);
- end
- if handles.cellselectionmatrix (row,col) > 2
- t = text (row, col, handles.labellist (i), 'VerticalAlignment', 'top', 'FontSize', 10);
- end
- hold on;
- end
-
- % Show the map in (row, column) format with row 1 at the top left
- axis ij;
- % Label axes
- axis ([1 handles.nrofrows 1 handles.nrofcols]);
- view (90, 90);
- % Interpolated
- shading interp;
- if handles.plotformat == 'M'
- handles.multipleplotnr = handles.multipleplotnr + 1;
- end;
- guidata (gcbo, handles);
-
- % -----

- function varargout = radiobutton23_Callback(h, eventdata, handles, varargin)
- % Umatrix
-
- % Initialize u-matrix
- for i = 1:handles.nrofrows
- for j = handles.nrofcols
- umatrix (i,j) = 0;
- end
- end
-
- %Calculate value for each cell of umatrix
- for i = 1:handles.nrofrows
- for j = 1:handles.nrofcols
- largest = 0;
- % Get the reference vector for the current unit using offset, as below. Distances from this vector will be calculated
- currentvectorindex = (handles.nrofcols * (i - 1)) + j;
- currentvector = handles.connections (currentvectorindex,:);
- % Get immediate neighborhood of current unit, with allowance for corners and edges. Immediate neighborhood
- % is a square of distance 1 around the current unit, ie, immediate side and diagonal units
- startrow = i - 1;
- if startrow < 1
- startrow = 1;
- end
- endrow = i + 1;
- if endrow > handles.nrofrows
- endrow = handles.nrofrows;
- end
- startcol = j - 1;
- if startcol < 1
- startcol = 1;
- end
- endcol = j + 1;
- if endcol > handles.nrofcols
- endcol = handles.nrofcols;
- end
- % For units at corners and edges, the number of neighbors is fewer than for internal units, and so the sum of distances
- % will be based on fewer values. This could skew the map, so a counter is used to normalize for this, as below
- counter = 0;
- % Get the sum of distances from the current unit to immediate neighbors, using the neighborhood spec derived above
- sumofdistances = 0;
- for k = startrow:endrow
- for m = startcol:endcol
- if ~((k == i) & (m == j)) % Don't count the distance of the current unit to itself, though it doesn't really matter, ie, should be 0

- % Get the connection vector of the neighboring unit currently in question
- adjoiningvectorindex = (handles.nrofcols * (k - 1)) + m;
- adjoiningvector = handles.connections (adjoiningvectorindex,:);
- % Add the euclidean distance between the connection vectors associated with the current unit and the current neighbor unit
- sumofdistances = sumofdistances + euclidean distance (currentvector, adjoiningvector);
- % Counter as above
- counter = counter + 1;
- end
- end
- end
- % Save the sum of distances in the umatrix, with normalization as above
- umatrix (i,j) = sumofdistances / counter;
- end
- end
-
- handles.cellselectionmatrix = zeros (handles.nrofrows, handles.nrofcols);
- % For each vector in turn
- for i = 1:handles.nrofdatavectors
- % Get the current vector
- smallest = 100000;
- currentvector = handles.vectorlist (i,:);
- % Generate a map for the current vector
- for j = 1:handles.nrofrows
- for k = 1:handles.nrofcols
- connectionindex = (handles.nrofcols * (j - 1)) + k;
- connectionvector = handles.connections (connectionindex,:);
- distance = euclidean distance (currentvector, connectionvector);
- if distance < smallest
- row = j;
- col = k;
- smallest = distance;
- end
- end
- end
-
- %Note that a vector is assigned to this cell
- handles.cellselectionmatrix (row,col) = handles.cellselectionmatrix (row,col) + 1;
-
- if handles.plotformat == 'M'
- subplot (handles.multipleplotrows, handles.multipleplotcols, handles.multipleplotnr);
- end
- surf (umatrix);
- if handles.cellselectionmatrix (row,col) == 1
- t = text (row, col, handles.labellist (i), 'VerticalAlignment', 'bottom', 'FontSize', 10);
- end

- if handles.cellselectionmatrix (row,col) == 2
- t = text (row, col, handles.labellist (i), 'VerticalAlignment', 'middle', 'FontSize', 10);
- end
- if handles.cellselectionmatrix (row,col) > 2
- t = text (row, col, handles.labellist (i), 'VerticalAlignment', 'top', 'FontSize', 10);
- end
- hold on;
- end
-
- %t = text (row, col, handles.labellist (i));
- % Show the map in (row, column) format with row 1 at the top left
- axis ij;
- % Label axes
- axis ([1 handles.nrofrows 1 handles.nrofcols]);
- axis off;
- view (90,90);
- % Interpolated
- shading interp;
- if handles.plotformat == 'M'
- handles.multipleplotnr = handles.multipleplotnr + 1;
- end;
- guidata (gcbo, handles);
-
-
-
- % PASSIVE TEXT BOXES
-
- % -----
- function varargout = edit13_Callback(h, eventdata, handles, varargin)
-
- % -----
- function varargout = edit14_Callback(h, eventdata, handles, varargin)
-
- % -----
- function varargout = edit15_Callback(h, eventdata, handles, varargin)
-
- % -----
- function varargout = edit16_Callback(h, eventdata, handles, varargin)
-
- % -----
- function varargout = edit18_Callback(h, eventdata, handles, varargin)
-
- % -----
- function varargout = edit22_Callback(h, eventdata, handles, varargin)
-
- % -----
- function varargout = edit23_Callback(h, eventdata, handles, varargin)
-
- % -----

- function varargout = edit24_Callback(h, eventdata, handles, varargin)
-
- % -----
- function varargout = edit25_Callback(h, eventdata, handles, varargin)
-
- % -----
- function varargout = edit26_Callback(h, eventdata, handles, varargin)
-
- % -----
- function varargout = edit27_Callback(h, eventdata, handles, varargin)
-
- % -----
- function varargout = edit28_Callback(h, eventdata, handles, varargin)
-
- % -----
- function varargout = edit29_Callback(h, eventdata, handles, varargin)
-
- % -----
- function varargout = edit30_Callback(h, eventdata, handles, varargin)
-
- % -----
- function edit35_CreateFcn(hObject, eventdata, handles)
- function edit35_Callback(hObject, eventdata, handles)

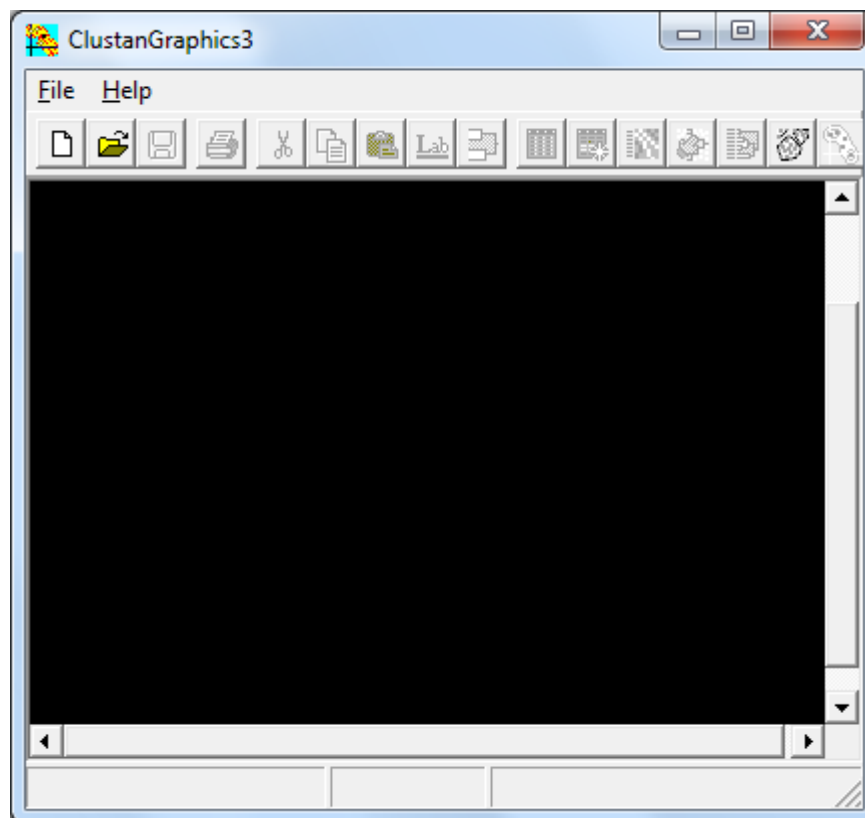
• 2.5 Hierarchical

- function f = clusterhierarchicaleuclid (m, llabels)
-
- [nrofrows nrofcols] = size(m);
-
- distancevector = pdist(m);
- tree = linkage(distancevector, 'method');
- dendrogram(tree, 0, 'labels', llabels, 'orientation','right');
- cophenetic = cophenet(tree,distancevector);
- f = cophenetic
-
- where:
- *m* is a data matrix
- *llabels* is a list of labels for the rows of *m*
- *pdist* creates the data matrix format required by Matlab
- *linkage* creates the cluster tree structure; the 'method' parameter specifies which type of linkage is required, ie, single, complete, etc.
- *dendrogram* constructs the cluster tree diagram
- *cophenetic* calculates the cophenetic correlation coefficient

Appendix 10: Cluster analysis version ClustanGraphics3:

ClustanGraphics3 is a program used for hierarchical cluster analysis developed by ClustanGraphics Inc.Ltd. It can display shaded representations of proximity matrices, dendrograms and scatterplots for 11 clustering methods: Single linkage, Complete linkage, Average linkage, Weighted Average linkage, Mean Proximity linkage, Centroid linkage, Median linkage, Increase in Sum of Squares (Ward's Method), Sum of Squares linkage, Flexible beta linkage, and Density search linkage. It can also provide a range of proximity measures, which differ according to the type of data: Euclidean distance, Squared Euclidean distance, Euclidean Sum of Squares, Jukes- Canto Genetic distance, Product-moment correlation, Pearson distance, and Jaccard similarity coefficient.

The user interface looks like this:



Here are the basic steps for using ClustanGraphics3:

- Click on file, then choose data matrix.

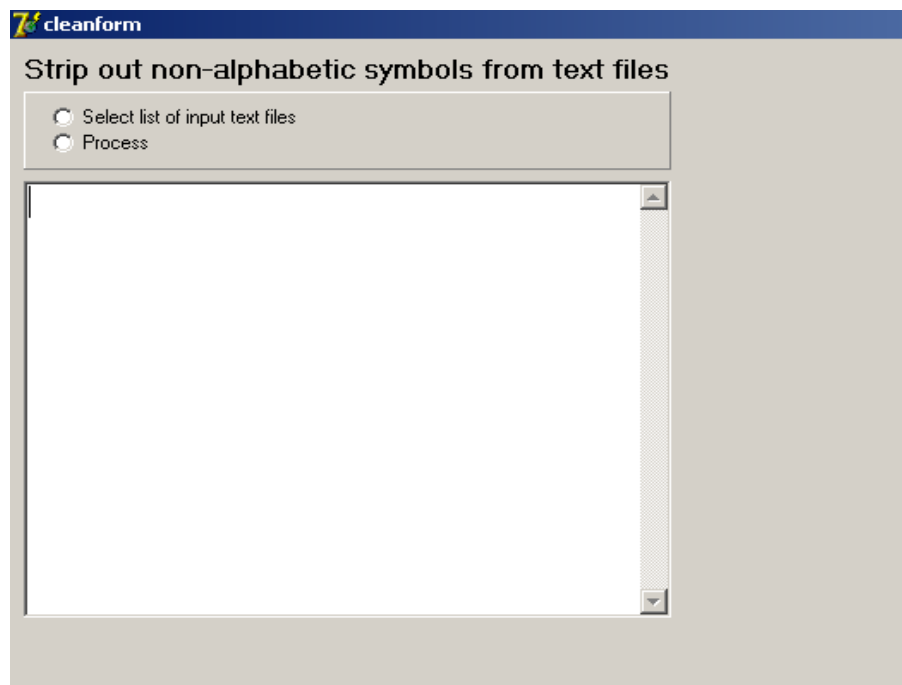
- Click on the distance to select the required distance between cases or clusters.
- Click on the required hierarchical clustering method.

Since the computational requirements of the current study was a very application-specific, it was necessary and more convenient to write some application-specific programmes than to search for suitable commercial, or use the existing, ones for research purposes: data pre-processing, data generation, and data interpretation. These were written by the researcher, jointly with Dr. Herman Moisl, in the School of English Literature, Language and Linguistics.

Appendix 11: Clean texts:

Cleantextfiles form 7 is a programme that removes chapter or section numbers and page references from the texts used in the study and as illustrated in Section (3.1) and shown in Table (3.2) in the forgoing discussion.

The user interface looks like this:



Here are the basic steps for using Clean texts:

- Manually edit out any material we don't want to include BEFORE cleaning, i.e. title, bibliographic information, chapter headings, page references etc.

- Make a list of file names, every file name must be EXACTLY the same as the name of the corresponding text document. For example, 'filenames.txt' contains one name, 'minstrel.txt', and this is the same as the name of the text file 'minstrel.txt'.
- Select 'select list of input text files' to load text file name list of the NN text files in a corpus.
- Click on 'process' to process and clean all the texts. The output prefixes 'cl' will be attached to the name of each file to distinguish it from the original: for example, the output for 'minstrel.txt' is 'clminstrel.txt'. The cleaned texts will be saved automatically.

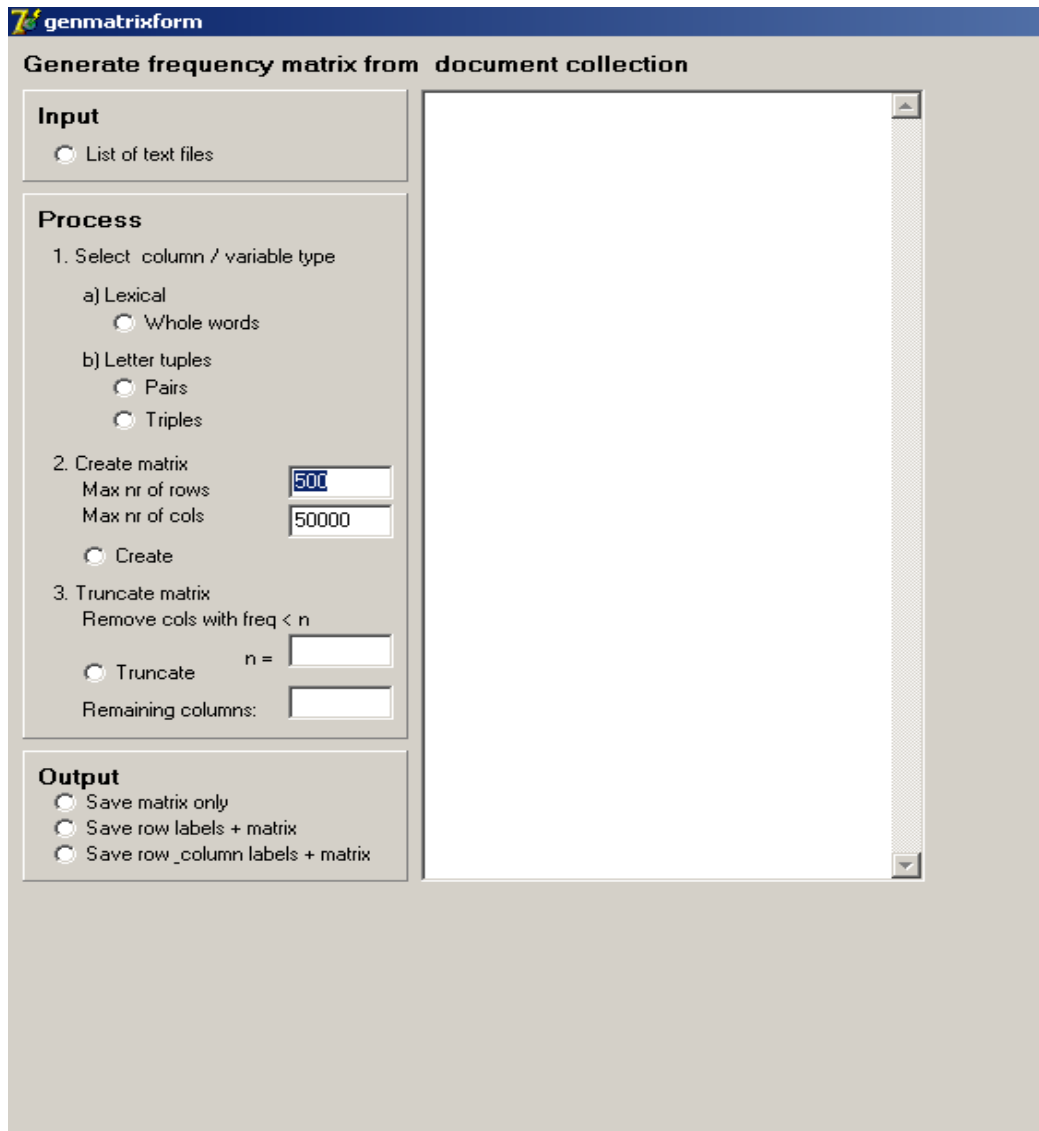
Appendix 12: Generate frequency matrix:

This programme is used to generate a frequency matrix. Given a list of textfile names and a corresponding set of texts such that, for $i = 1..363$, $name_i$ denotes $text_i$, generate a vector for each $text_i$, where:

- Each $text_i$ has a column vector representing a function word, with the result that the matrix has as many columns n as there are function words in the corpus.
- The value in each row vector j , for $j = 1..n$, is the number of times the associated function word occurs in $text_i$.

For convenience, a set of n text vectors is represented as a matrix Mn in which the rows represent texts and the columns represent function words.

The user interface looks like this:



Here are the basic steps for using Generate Frequency matrix:

1. Click on the 'Input' box, 'text file list' to upload a list of text filenames.
2. Click 'process' box, then click 'lexical', and select button 'a' (whole words).
3. Click 'Create matrix' box, then enter the maximum number of columns and rows required for the generation of the data matrix.
4. Click 'Create' to generate lexical type frequency matrix M .
5. As required, click 'truncate matrix box', then enter the number of columns one wants to remove and click 'truncate'. A number will appear to the user during execution referring to the remaining columns in the data matrix.
6. Click the 'Output' box, the sequence of radio buttons allows M to be formatted in various ways:

- 'Save matrix only' saves the matrix without row or column labels.
 - 'Save row labels /matrix' saves the number of rows labels with the matrix.
 - 'Save row/column labels /matrix' saves the number of rows and columns labels with the matrix in ways indicated by the button options.
7. The text space on the right of the interface shows messages to the user during program execution and also allows display of various kinds of interim output during program debugging.

Appendix 13: Edit matrix:

This programme is used to edit a frequency matrix M generated by GenerateFreqMatrix just described above. Given M as input data matrix, EditMatrix allows:

- i. Normalization for variation in text file length
- ii. Several types of dimensionality reduction, including:
 - Removal of frequency- N columns and columns associated with function words.
 - Retention of explicitness-specified columns.
 - Removal / retention of columns on the basis of column standard deviation.
 - Removal / retention of columns on the basis of column term frequency / inverse document frequency.
 - Removal / retention of columns on the basis of column Poisson distribution.
 - Removal / retention of columns on the basis of column entropy.
- iii. Various two dimensionality reduction features:
 - Sorting of rows and columns by frequency and covariance.
 - Calculation, sorting, and output of covariance and correlation matrices.
- iv. Output of the edited matrix in a variety of formats.

The interface for Editmatrix looks like this:

The screenshot displays the 'Edit matrix' software interface, organized into several functional panels:

- Input:** Options for matrix file attributes (Nr of rows/columns header, Row labels, Column labels) and matrix reading.
- Document length normalization:** Methods such as Log, Maximum term frequency within document, Mean term frequency within document, Mean document length across collection, Cosine, and Relative frequency.
- Weighting:**
 - Variance / standard deviation:* Weight column vectors by variance or standard deviation.
 - TF.IDF:* Weight column vectors by TF.IDF.
 - Poisson term distribution:* Weight column vectors by Poisson index.
 - Noise / signal ratio:* Weight column vectors by signal.
- Utilities:**
 - Sort:* Sort rows or columns by frequency or variance, or sort rows by column magnitudes ascending.
 - Select:* Select first n rows or columns.
 - Covariance / correlation matrices:* Calculate and save covariance or correlation matrices, output sorted descending, or output diagonal.
- Dimensionality reduction:**
 - Heuristics:* Remove all columns with frequency ≤ 1 or remove function word columns.
 - Retain columns selected by keyword:* Open keyword list or apply.
 - Variance / standard deviation:* Remove columns with std dev $< n$ or retain n highest-std dev columns.
 - TF.IDF:* Remove columns with TF.IDF $< n$ or retain n highest-TF.IDF columns.
 - Poisson term distribution:* Remove cols w/mean-variance $< n$ or retain n highest mean-variance cols.
 - Noise / signal ratio:* Remove cols w/signal $< n$ or retain n highest-signal cols.
 - Linear Principal Components:* Preprocessing (Mean-centre matrix columns) and Import eigen-matrices (Axis length of eigenmatrices, Import eigenvalue matrix, Import eigenvector matrix).
- Current matrix dimensions:** Fields for Rows and Cols.
- Output:** Option to save row / column header.

Here are the basic steps for using Edit matrix in ways that are relevant to the present study:

For Normalization and Dimensionality reduction:

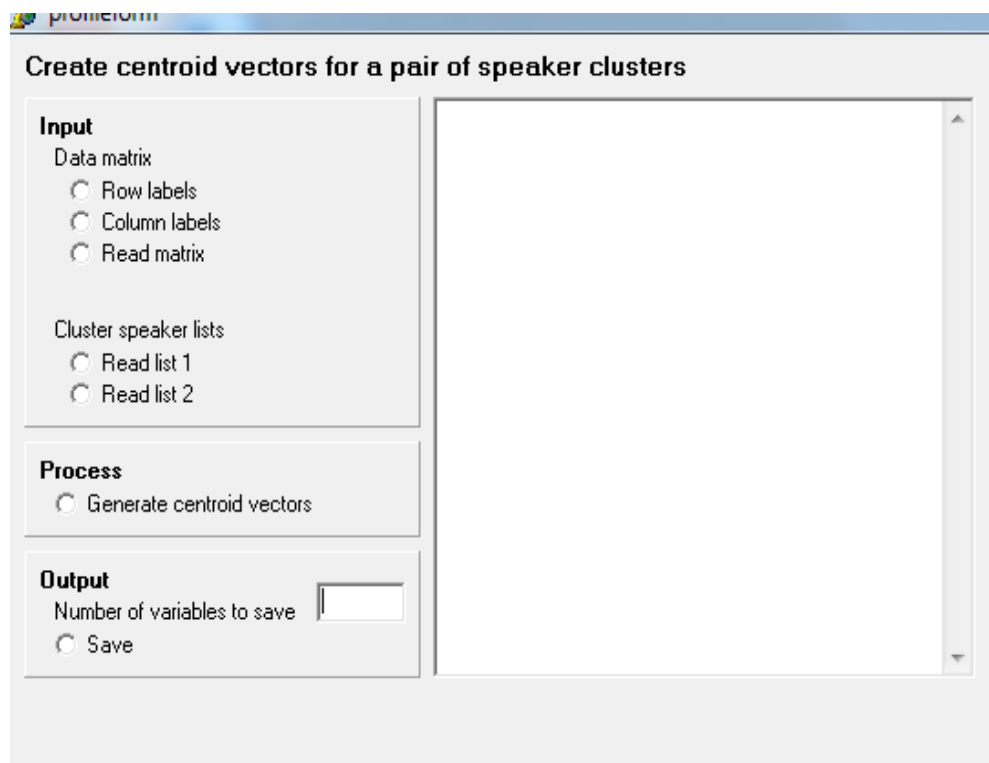
- Upload an input data matrix
- Click on the required method to compensate for variation in text file length or to reduce dimensionality and extract the the most important variables.

- Click on any button in the Utilities to sort the rows and columns of the data matrix.
- Save the resulting matrix.
- Finally, the text box on the right of the interface shows messages to the user during program execution and also allows display of various kinds of interim output during program debugging.

Appendix 14: Centroid vectors:

This programme is used to calculate centroid vectors for cases or clusters and compare them by taking the centroid in each text file vector that constitute clusters with a particular emphasis on identifying the variables most important in distinguishing the textfiles in each cluster.

The interface for Editmatrix looks like this:



The steps for using this programme are:

- Click 'Row labels', 'Column labels', and 'Read matrix' in succession.
- Click 'Read list1' and 'Read list2' to load the involved file name clusters.

- Click 'Create centroid vectors' and wait until the completion message appears in the box.
- Click 'enter' to put the number of variables one wants to save.
- Click 'Save' to save the generated centroids.
- Use the resulting centroids to create a bar plot with SPSS or MATLAB.

Bibliography

In Humanities, the most common formulated referencing system is called the Harvard system. There is, however, no definitive version of the Harvard system. Most universities compiled and edited their own versions. The 'Style Guide for work in Language and Linguistics 2014-15: 22', i.e. the Newcastle Harvard, edited for the school of English Literature, Language and Linguistics by Heike Pichler is used here.

Abbasi, A., and Chen, H. 2005. 'Applying authorship analysis to extremist-group web forum messages'. *IEEE Intelligent Systems* 20: 67-75.

Aggarwal, C., Hinneburg, A., and Keim, D. 2001. 'On the surprising behaviour of distance metrics in high dimensional space'. Last accessed 11 August 2010, from: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.23.7409&rep=rep1&type=pdf>.

Aleamoni, L. M. 1976. 'The relation of sample size to the number of variables in using factor analysis'. *Educational and Psychological Measurement* 36: 879-883.

Allen Michael Patrick. 1997. *Understanding Regression Analysis*. USA: A Division of Plenum Publishing.

Allinson, N., Yin, H., Allinson, L., Slack, J. (eds). 2001. *Advances in Self-Organizing Maps*. Berlin: Springer-Verlag Ltd.

Anderberg, M. R. 1973. *Cluster Analysis for Applications*. London: Academic Press, Inc.

Andrewes J. L and McNicholas P. D. 2013. 'Variable Selection for Clustering and Classification'. Last accessed 22 May 2013, from: <http://arxiv.org/abs/1303.5294v1>.

Andrewes J. L. 2014. 'Variable Selection for Clustering and Classification'. *Journal of Classification* 13: 136-153.

- Annas Suwardi, Takenori Kanai and Shuhei Koyama. 2007. 'Principal Components Analysis and Self-Organizing Map for Visualizing and Classifying Five Risks in Forest Regions'. *Japanese Society of Agricultural Informatics* 16: 44-51.
- Anster, J. 1835. *Faustus. A Dramatic Mystery*. London: Longman Press.
- Arabie, P., Hubert, L. and de Soete, G. (eds).1992. *Clustering and Classification*. New Jersey: World Scientific Press.
- Argamon-Engelson, S., Koppel, M., and Avneri, G. 1998. 'Style-based text categorization: What newspaper am I reading?'. Last accessed 23 May 2010, from: <https://www.aai.org/Papers/Workshops/1998/WS-98-05/WS98-05-001.pdf>.
- Argamon, S., Koppel, M., and Avneri, G. 1998. 'Routing documents according to style'. Last accessed 11 July 2009, from: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.52.688&rep=rep1&type=pdf>.
- Argamon, S. and Levitan, S. 2005. 'Measuring the usefulness of function words for authorship attribution'. Last accessed 11 July 2009, from: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.71.6935&rep=rep1&type=pdf>.
- Arppe, A. 2008. 'Univariate, bivariate, and multivariate methods in corpus-based lexicography: A study of synonym'. *HELDA* 44: 1-600.
- Ashton, R. D. 1977. 'Coleridge and Faust'. *Review of English Studies* 28: 156-167.
- Baayen, H., van Halteran, H., Neijt, A., Tweedie, F. 2002. 'An Experiment in Authorship Attribution'. Last accessed 11 July 2009, from: <http://www.sfs.uni-tuebingen.de/~hbaayen/publications/BaayenVanHalterenNeijtTweedieJADT2002.pdf>.
- Baayen R. H. 1996. 'Using Syntactic Annotation to Enhance Authorship Attribution'. *Literary and linguistic computing* 11: 121-131.

-2001. *Word Frequency Distributions*. Dordrecht: Kluwer.
-2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.
- Bailey R. W. 1979. 'The future of computational stylistics'. *Association for Literary and Linguistic Computing* 2: 61-70.
- Baker, F. and Hubert, L. 1975. 'Measuring the Power of hierarchical cluster analysis'. *Journal of the American Statistical Association* 70: 31-38.
- Baker Kirk. 2013. 'Singular Value Decomposition Tutorial'. Last accessed 20 March 2014, from: http://webcache.googleusercontent.com/search?q=cache:pAPZQTu-ldYJ:https://www.ling.ohiostate.edu/~kbaker/pubs/Singular_Value_Decomposition_Tutorial.pdf+&cd=1&hl=en&ct=clnk&gl=uk.
- Balakrishnan, N., Voinov, V., Nikulin, M. 2013. *Chi-Squared Goodness of Fit Tests with Applications*. USA: Elsevier Academic Press.
- Bartke, K. 2005. '2D, 3D and high-dimensional data and information visualization'. Last accessed 5 June 2013, from: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.83.3421&rep=rep1&type=pdf>.
- Barret, P.T. and Kline, P. 1981. 'The observation to variable ratio in factor analysis'. *Personality Study and group behaviour* 1: 23-33.
- Bauer, I. 2015a. 'Aus Versus Von'. Last accessed 26 August 2015, from: <http://german.about.com/od/grammar/a/Aus-Versus-Von.htm>.
- Bauer, I. 2015b. 'German Preposition 'Aus''. Last accessed 26 August 2015, from: <http://german.about.com/od/grammar/a/German-Preposition-Aus.htm>.
- Bauer, I. 2015c. 'Dative Prepositions'. Last accessed 26 August 2015, from: <http://german.about.com/od/grammar/ht/Dative-Prepositions.htm>.

- Beach, Gleneden. 2002. 'Hierarchical clustering'. In McCune, B. and Grace, J.B (eds) *Analysis of Ecological Communities*. Mjrm Software Design. 86-96.
- Bee, R. E. 1970. 'Statistical Methods in the Study of the Masoretic text of the Old Testament'. *Journal of the Royal Statistical Society* 134: 611-622.
- Behrens, J.T. 1997. 'Principles and Procedures of exploratory data analysis'. *American Psychological Association* 2: 131-160.
- Belew, R. 2000. *Finding Out About: A Cognitive Perspective in Search Engine Technology and the WWW*. Cambridge: Cambridge University Press.
- Bell Alan, Jason, M. Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky. 2009. 'Predictability Effects on Durations of Content and Function Words in Conversational English'. *Journal of Memory and Language* 60: 92-111.
- Bell, E. J., Berridge, D., and Rayson, P. 2009. 'Measuring style with the authorship ratio: An invariant metric of lexical similarity'. *Corpus Linguistics* 20: 1-14.
- Bellman, R. 1961. *Adaptive Control Processes: A Guided Tour*. USA: Princeton University Press.
- Ben-Hur, A. Elisseeff, A., and Guyon, I. 2002. 'A Stability based method for discovering structure in clustered data'. Last accessed 20 July 2009, from: <http://psb.stanford.edu/psb-online/proceedings/psb02/benhur.pdf>.
- Bennett, W.R. 1976. *Scientific and Engineering Problem-Solving with the Computer*. Englewood Cliffs, NJ: Prentice Hall, Inc.
- Berry William D. and Feldman Stanley. 1985. *Multiple Regression in Practice*. London: SAGE Publications Ltd.
- Best, K-H. 1997. 'The distribution of word and sentence length'. *Glottometrika* 16:152:162.
-2001. *Häufigkeitsverteilungen in Texten*. Göttingen: Peust and Gutschmidt Verlag.

- Beyer, K. et al. 1999. 'When is Nearest Neighbor Meaningful?'. Last accessed 20 July 2009, from: <http://www.loria.fr/berger/Enseignement/Master2/Exposes/beyer.pdf>.
- Bhattacharjee Anol. 2012. *Social Science Research: Principles, Methods, and Practices*. USA: Global Text Project.
- Binongo, J.N.G. 2003. 'Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution'. *Chance* 16: 9-17.
- Bishop, C. M. 1995. *Neural Networks for Pattern Recognition*. USA: Oxford University Press.
- Bissell A F. 1995. 'Weighted cumulative sums for text analysis using word counts'. *Journal of the Royal Statistical Society* 158: 525-545.
- Bode, C. 2008. 'Faustus From the German of Goethe Translated by Samuel Taylor Coleridge. Coleridge as translator of Faust'. Last accessed 18 April 2008, from: <http://www.friendsof Coleridge.com/Faustus.htm>.
- Borg, I and Groenen P. 1997. *Modern Multidimensional Scaling: Theory and Applications*. New York: Springer.
- Bornheimer, B., Fitzpatrick, R., Lehmann, S., Pierce, M. and Whalen, M. 2008. 'Reliability and validity in research'. Last accessed 10 February 2009, from: [https:// www.docs.ReliabilityValidity.Courses.htm](https://www.docs.ReliabilityValidity.Courses.htm).
- Boyle, N. 1987. *Goethe Faust. Part One*. USA: Oxford University Press.
- Boyle, N. and Guthrie, J. 2002. *Goethe and the English Speaking World*. USA: Camden House.
- Bozkurt, I., N., Baghoglu, O., and Uyar, E. 2007. 'Authorship Attribution'. *Computer and information sciences* 1-5.
- Bray, E. 2007. 'Sex and Drugs and English Literature: Coleridge and a Faustian Pact'. Last accessed 8 April, 2008, from: <http://www.independent.co.uk/news/uk/literature-coleridge.htm>.

- Brennan, M. and Greenstadt, R. 2009. 'Privacy and Stylometry: Practical attack against authorship recognition techniques'. Last accessed 10 May 2010, from: https://www.cs.drexel.edu/~greenie/brennan_paper.pdf.
- Brett, R. L. and Jones, A. R. 1963. *Wordsworth and Coleridge Lyrical Ballads 1798-1800: The Text of the 1798 Edition with Additional 1800 Poems and Prefaces*. London: Methuen and Co Ltd.
- Brinegar, C.S. 1963. 'Mark Twain and the Quintus Curtius Snodgrass letters: A Statistical test of authorship'. *Journal of the American Statistical Association* 58: 85-96.
- Burrowes, B. 2007. 'Finding Faustus'. Last accessed 10 April 2009, from: <http://archive.umt.edu/montanan/f07/faustus.php>.
- Burrows, J. F. 1987. 'Computation into criticism: A study of Jane Austen's novels and an experiment in method'. *The Journal of English and Germanic Philology* 88: 111-114.
- Burrows, J. F. 1992. 'Computers and the study of literature'. In Butler, C. S. (ed.) *Computers and Written Texts*. Oxford: Blackwell. 167-204.
- Burrows, J. F. 2002a. 'Delta: A measure of Stylistic difference and a guide to likely authorship'. *Literary and Linguistic Computing* 17: 267-287.
-200b. 'The Englishing of Juvenal: Computational Stylistics and Translated Texts'. *Style* 36: 677-94.
-2003. 'Textual analysis'. In Schreibman Susan, Siemens Ray, and Unsworth John (eds) *A companion to Digital Humanities*. Oxford: Blackwell. 323-347.
- Burwick, F. 2008a. 'On Coleridge as translator of Faustus from the German of Goethe'. *The European Romantic Review* 19: 247-252.

-2008b. 'A gentleman of literary eminence. *Coleridge as translator of Faust*'. Last accessed 18 April 2008, from: Coleridge <http://www.friendsofColeridge.com/Faustus.htm>.
-2008c. 'An orphic tale: Goethe's Faust translated by Coleridge'. In Fitzsimmons, L. (ed.) *International Faust Studies: Adaptation, Reception, Transition*. London: Continuum. 124-145.
- Burwick, F. and McKusick, J. 2007. *Faustus from the German of Goethe Translated by Samuel Taylor Coleridge*. Great Britain: Oxford University Press.
- Camastra, F. 2003. 'Data dimensionality estimation methods: A survey'. *Pattern Recognition* 36: 2945-54.
- Canter, D. 1992. 'An Evaluation of the "Cusum" Stylistic Analysis of Confessions'. *Expert Evidence* 1: 93-99.
- Casey Paul Foley. 1981. 'John Martin Anster: His Contributions to Anglo-German Studies'. *MLN* 96: 654-665.
- Chalmers, A. F. 1999. *What is this thing called Science?*. 3rd ed. UK: Open University Press.
- Chan Winnie Wing Yi. 2006. 'A Survey on Multivariate Data Visualization'. *Lecture Notes in Computer Science* 6771: 221-231.
- Charles Sander Peirce. 2014. 'Stanford Encyclopaedia of Philosophy'. Last accessed 13 June 2014, from: <http://plato.stanford.edu/entries/peirce/>.
- Chaska, C. 2001. 'Empirical evaluations of language-based author identification techniques'. *Forensic Linguistics* 81: 1-65.
-2005. 'Who's at the keyboard? Authorship attribution in digital evidence investigations'. *International Journal of Digital Evidence* 4: 1-13.
- Chen C. H. 2012. 'Feature Selection for Clustering by Exploring Nearest and Farthest Neighbours'. *Information Sciences* 318: 1-156.

- Chen Zhili, Liusheng Huang, Wei Yang, Peng Meng, and Haibo Miao. 2012. 'More than Word Frequencies: Authorship Attribution via Natural Frequency Zoned Word Distribution Analysis'. Last accessed 20 May 2014, from: <http://dblp.uni-trier.de/db/journals/corr/corr1208.html#abs-1208-3001>.
- Chung Cindy and Pennebaker. 2007. 'The Psychological Functions of Function Words'. In K. Fiedler (ed.) *Social Communication*. New York. 343-359.
- Cios, K. J., Pedrycz, Swiniarski, R., and Kurgan, L. A. 2007. *Data Mining: A Knowledge Discovery Approach*. Canada: Springer Science Bussiness Media.
- Clarke, R. Resson, HW, Wang, A., Xuan, J., Liu M. C., Gehan, E. A., and Wang, Y. 2008. 'The Properties of high dimensional data spaces: implications for exploring gene and protein expression data'. *NCBI* 8: 37-49.
- Clarke, G. M. and Cooke, D. 1998. *A Basic Course in Statistics*. UK: John Wiley & Sons.
- Classe, O. 2000. *Encyclopedia of Literary Translation into English*. USA: Fitzroy Dearborn Publishers.
- Clement, R. and Sharp, D. 2003. 'N-gram and Bayesian classification of documents'. *Literary and Linguistic Computing* 18: 423-447.
- Coleridge, S.T. 1817. *The Sibylline Leaves: A Collection of Poems*. Curtis printer, Camberwell, London.
-1912. *The Complete Poetical Works of STC including Poems and Versions of Poems NOW Published for the First Time in Two Volumes. Volume I Poems. Volume II Dramatic Works and Appendices with Textual and Biographical Notes*. Great Britain: Oxford University Press.
- Coleridge, H. N. 1835. *The Specimens of the Table Talk of the Late S. T. Coleridge in Two Volumes*. New York: Harper and Brothers.
-1844. *The Poetical and Dramatic Works of S. T. Coleridge*. London: William Pickering.

-1895. *The Poetical Works of S. T. Coleridge: Reprinted from the Early Editions with Memoir, Notes, etc.* London: Frederick Warne and Co. and New York.
- Constantine, D. 2005. *Faust, Part I*. USA: A. S. Byall (Penguin Classics).
-2006. 'Literary translation: German'. In France, P. and Haynes, K. (eds) *The Oxford History of Literary Translation in English*. Oxford University Press. 211-229.
- Cook, J. S. Homepage. 2009. 'Lecture notes for linear Algebra'. Last accessed 12 July 2013, from: <http://www.supermath.info/math321.pdf>
- Cottrell, M., de Bodt, E., and Verleysen, M. 2001. 'A Statistical tool to assess the reliability of self-organizing maps'. In Allinson, N., Yin, H., Allinson, L., Slack, J. (eds) *Advances in Self-Organizing Maps*. Berlin: Springer. 7-14.
- Coyotl-Morales, Luis Villasenor-Pineda, Manuel Montes-y- Gomez, and Paolo Rosso. (2006). 'Authorship Attribution Using Word Sequences'. Last accessed 20 July 2009, from: http://users.dsic.upv.es/~proso/resources/CoyotlEtAl_CIARP06.pdf.
- Crane, M. T. 2001. *Shakespeare's Brain: Reading with Cognitive Theory*. Princeton, NJ: Princeton University Press.
- Craig, H. 1999. 'Authorial attribution and computational stylistics: if you can tell authors apart, have you learned anything about them?'. *Literary and Linguistic Computing* 14: 103-113.
-2003. 'Stylistic analysis and authorship studies'. In Schreibman, S., Siemens, R., and Unsworth, J. (eds) *A companion to Digital Humanities*. Oxford: Blackwell. 271-288.
-2008. 'The stylometric analysis of Faustus, from the German of Goethe'. *The Journal of the Friends of Coleridge* 32: 85-88.

- Crick, J. 2008. 'Faustus from the German of Goethe translated by Samuel Taylor Coleridge. Oxford University Press edited by Frederick Burwick and James C. McKusick'. *The Journal of the Friends of Coleridge* 32: 71-84.
- Crystal, D. and Davy, D. 1969. *Investigating English Style*. London: Longman.
- Culpeper, J. 2002. 'Computers, language and characterization: An analysis of six characters in Romeo and Juliet'. In Melander-Marttala, U., Ostman, C. and Kyto, M. (eds) *Conversation in Life and in Literature: Papers from the ASLA symposium*. Uppsala: Universitetsstryckeriet. 11-30.
- Cyran, A. K. and Stanczyk, U. 2007. 'Machine learning approach to authorship attribution of literary texts'. *International Journal of Applied Mathematics and Informatics* 4: 151-158.
- Dabagh, R. M. 2007. 'Authorship attribution and statistical text analysis'. *Metodološki zvezki* 2: 149-163.
- Dale, R., Moisl, H. and Somers, H. 2000. *Handbook of Natural Language Processing*. New York: Marcel Dekker.
- Damerau Fred, J. 1975. 'The Use of Function Word Frequencies as Indicator of Style'. *Computers and Humanities* 19: 271-280.
- Datta, K. B. 2004. *Matrix and Linear Algebra*. New Delhi: India private Limited.
- De Paulo, P. 2011. 'Sample Size for Qualitative Research'. *NCBI* 18: 179-183.
- DeSilva, V. and Tenebaum, J. 2003. 'Unsupervised Learning of Curved Manifolds'. In David D. Denison, Mark H. Hansen, Christopher C. Holmes, Bani Mallick, Bin Yu (eds) *Nonlinear Estimation and Classification*. Berlin: Springer.
- De Vel, O., Anderson, A., Corney, M., and Mohay, G. M. 2001. 'Mining e-mail content for author identification forensics'. *ACMDL* 30: 55-64.
- Dewell, R. B. 2015. *The Semantics of German Verb Prefixes*. The Netherlands: John Benjamins Publishing Company.

- Deza, M. and Deza, E. 2009. *Encyclopedia of Distances*. Berlin: Springer.
- Diederich, J., Kindermann, J., Leopold, E., and Paass, G. 2000. 'Authorship attribution with Support Vector Machines'. *Applied Intelligence* 19: 109-123.
- Dikken, M. D. and Tortora, C. 2005. *The Function of Function Words and Functional Categories*. The Netherlands: John Benjamins Publishing.
- Dixon, P. and Mannion, D. 1993. 'Goldsmith's Periodical Essays: a Statistical Analysis of Eleven Doubtful Cases'. *Literary and Linguistic Computing* 8: 1-19.
- Donald J. Lewis and Wilfred Kaplan. 2007. *Calculus and Linear Algebra. Vol 2: Vector Spaces, Many Variables Calculus, and Different Equations*. Michigan: University of Michigan Library.
- Duda, R., Hart, P. and Stork, D. 2001. *Pattern Classification*. 2nd ed. New York: John Wiley and Sons.
- Dumais, S., J. Platt, J., Heckerman, D. and Sahami, M. 1998. 'Inductive learning algorithms and representations for text categorization'. *ACMDL* 148-52.
- Dunning, T. 1994. 'Statistical Identification of Language'. *Technical Report MCCA* 273:1-29.
- Durrant, R. J. and Kaban, A. 2009. 'When is 'Nearest Neighbour' Meaningful: A Converse Theorem and Implications?'. *Journal of Complexity* 25: 385-397.
- Eddy, H. T. 1887. 'The characteristic curves of composition'. *Science* 297.
- Eder, M. 2011. 'Style-Markers in Authorship Attribution A Cross-Language Study of the Authorial Fingerprint'. *Studies in Polish Linguistics* 6: 99-114.
- Ehrenberg, A.S.C. 1982. *A Primer in Data Reduction: An Introductory Statistics Textbook*. England: John Wiley and Sons Ltd.
- Elderton, P. 1949. 'A few statistics on the length of English words'. In Grzybeck, P. (ed.) *Contributions to the Science of Text and Language. Word length studies and related issues*. The Netherlands: Springer. 15-90.

- Engell, J. 2012. 'Faustus: From the German of Goethe, Translated by Samuel Taylor Coleridge. Coleridge as translator of Faust'. Last accessed 18 April, 2008, from: <http://www.friendsofColeridge.com/Faustus.htm>.
- Enkvist, N. E. 1964. 'On defining Style'. In Enkvist, N. E, Spencer, J. and Gregory, M. (eds) *Linguistics and Style*. Oxford University Press.
- Everitt, B. S. and Dunn, G. 2001. *Applied Multivariate Data Analysis*. UK: John Wiley & Sons Ltd.
- Everitt, B. S., Landau, S., and Leese, M. 2001. *Cluster Analysis*. 4th ed. UK: John Wiley & Sons Ltd.
- Feiguina, O. and Hirst, G. 2007. 'Forensic authorship attribution for small texts'. Last accessed 12 August 2008, from: <http://ceur-ws.org/Vol-276/paper3.pdf>.
- Fenton, J. 2008. 'Faust Lost in Translation?'. Last accessed 12 April 2008, from: <http://www.guradian.co.uk/books/2008/apr/12/poetry.theatre>.
- Fogarassy, Ágnes Vathy and Abonyi, János. 2013. *Graph Based Clustering and Data Visualization Algorithms*. London: Springer-Verlag.
- Fitzsimmons Lorna. 2008. *International Faust Studies: Adaptation, Reception, Translation*. Great Britain: MPG Biddles Ltd, Kings Lynn, Norfolk.
- Forman G. 2003. 'An extensive empirical study of features selection metrics for text classification'. *Journal of machine learning research* 3: 1289- 1305.
- Forsyth, R.S. and Holmes, D.I. 1996. 'Feature-finding for text classification'. *Literary and Linguistic Computing* 11: 163-174.
- Forsyth, R. S., Holmes, D. I., and Tse, E. 1999. 'Cicero, Sigonio, and Burrows: Investigating the authenticity of the Consolatio'. *Literary and Linguistic Computing* 14: 375-400.
- Forsyth, R. S. 2007. 'Notes on authorship attribution and text classification'. Last accessed 14 July 2009, from: <http://www.cs.nott.ac.uk/~axc/DReSS/LFAS08.pdf>.

- Francois, D., Wertz, D., and Verleysen, M. 2007. 'The concentration of fractional distances'. *IEEE* 19: 873-86.
- Frank, I. E. and Todeschini, R. 1994. *The Data Analysis Handbook*. Netherlands: Elsevier science.
- Fucks, W. 1952. 'On mathematical analysis of style'. *Biometrika* 39: 122-129.
-1954. 'On Nahordnung and Fernordnung in samples of literary texts'. *Biometrika*: 41, 116-132.
- Fucks, W. and Lauter, J. 1965. 'Mathematische analyze des literarischen stils'. In Kreuzer, H. and Gunzenhausers, R (eds) *Mathematik und Dichtung*. Munich: Nymphenburger Verlagsbuckhandlung.
- Fung, G. 2003. 'The Disputed Federalist Papers: SVM Feature Selection via Concave Minimization'. *In proceedings of the 2003 Conference on Diversity in Computing* 42-46.
- Gaborski, R. S. 2014. 'Principal Components Analysis'. Last accessed 14 May 2014, from:http://www.cse.psu.edu/~rtc12/CSE586Spring/lectures/pcaLectureShort_6pp.pdf.
- Gauch, H. G. 2003. *Scientific Method in Practice*. Cambridge: Cambridge University Press.
-2012. *Scientific Method in Brief*. Cambridge: Cambridge University Press.
- Gan, G., C. Ma, and J. Wu. 2007. *Data Clustering. Theory, Algorithms, and Applications*. Alexandria VA: American Statistical Association.
- Gareth James, Witten Daniela, Hastie Trevor, Tibshirani Robert. 2013. *An Introduction to Statistical Learning with Applications in R*. Springer-Verlag New York.
- Garcia, A. M. and Martin, J. C. 2007. 'Function words in authorship attribution studies'. *Journal of Linguistic and Computing* 22: 49-66.

- George Dallas. 2013. 'Principal Components Analysis for Dummies: Eigenvectors, Eigenvalues, and Dimension Reduction'. Last accessed 10 April 2014 from: <https://georgemdallas.wordpress.com/2013/10/30/principal-component-analysis-4-dummies-eigenvectors-eigenvalues-and-dimension-reduction/>.
- Germano Tom. 1999. 'Self-Organizing Maps'. Last accessed 10 April 2014, from: <http://davis.wpi.edu/~matt/courses/soms/>.
- Glass Derek. 2005. *Goethe in English: A Bibliography of the Translations in the Twentieth Century*. England: Maney Publishing.
- Greenwood Priscilla, E. and Mikhail S. Nikulin. 1996. *A Guide to Chi-Squared Testing*. Canada: John Wiley & Sons Ltd.
- Gnanandesikan, R., Tsao, S., Kettenring, John. 1995. 'Weighting and selection of variables for cluster analysis'. *Journal of Classification* 12: 113- 136.
- Gnanadesikan, R. 1997. *Methods for Statistical Data Analysis of Multivariate Observations*. 2nd ed. New York: Wiley-Interscience.
- Gordon, A. 1992. 'Hierarchical classification'. In P. Arabie, L. Hubert and G. de Soete (eds) *Clustering and Classification*. New Jersey: World Scientific Press.
-1999. *Classification*. 2nd ed. London: Chapman & Hall.
- Gore, P. 2000. 'Cluster Analysis'. In Tinlsey, H. and Brown, S. (eds) *Handbook of Applied Multivariate Statistics and Mathematical Modelling*. New York: Academic Press. 297-321.
- Gower, J. C. 1985. 'Measures of similarity, dissimilarity, and distances'. In Kotz, N. L., Johnson, R. A., and Read, C. B. (eds) *Encyclopedia of Statistical Sciences*. New York: John Wiley and Sons. 397-405.
- Greenwood, H. H. 1995. 'Common word frequencies and authorship in Luke's Gospel and Acts'. *Literary and Linguistic Computing* 10: 183-187.

- Greenwood Priscilla E. and Mikhail S. Nikulin. 1996. *A Guide to Chi-Squared Testing*.
Canada: John Wiley and Sons Ltd.
- Green P. E., Carmone F. J., Smith S. M. 2011. *Cluster Analysis Revision of
Multidimensional Scaling, Section 5: Dimensionality Reducing Methods and
Cluster Analysis*. Addison Wesley.
- Grieve, J. W. 2005. *Quantitative Authorship Attribution: A History and an Evaluation of
Techniques*. Canada: Simon Fraser University.
-2007. 'Quantitative authorship attribution: an evaluation of techniques'.
Literary and linguistic Computing 22: 251-270.
- Groenen, P. and Van de Velden, M. 2005. 'Multidimensional Scaling'. In Everitt, B and
Howell, D. (eds) *Encyclopedia of Statistics in Behavioural Science*. Hoboken NJ:
Wiley. 1280-1289.
- Gross, J. and Yellen, J. 2006. *Graph Theory and its Applications*. 2nd ed. London:
Chapman and Hall.
- Grovier, K. 2008. 'Coleridge and Goethe, together at last'. Last accessed 10 June 2008,
from: https://login.the-tls.co.uk/?gotoUrl=http%3A%2F%2Fwww.the-tls.co.uk%2Ftls%2Freviews%2Fother_categories%2Farticle757956.ece.
- Grzybek, P. 2007. 'History and methodology of word length studies. The State of the
Art'. In Grzybeck, P. (ed.) *Contributions to the Science of Text and Language*.
Dordrecht, NL: Springer. 15-90.
- Grzybeck, P., Ernst, S., Kelih, E. 2007. 'The relation of word length and sentence
length: The inter-textual perspective'. In Decker, R. and Lenz, H. (eds) *Advances
in Data Analysis*. Berlin: Springer. 611-618.
- Grzybeck, P. 2014. 'The Emergence of Stylometry: Prolegomena to the History of Term
and Concept'. In Kroo, Katalin, Torop, Peeter (eds) *Text within Text. Culture
within Culture*. Budapest, Tarty: LHarmattan. 58-75.

- Guadagnoli, E. and Velicer, W.F. 1988. 'Relation of sample size to the stability of component patterns'. *Psychological Bulletin* 103: 265-275.
- Guiraud, H. 1954. *Les caracteres statistique du vocabulaire*. Paris: Presses Universitaires de France.
- Hair, J., Anderson, R., Tatham, R., and Black, W. 1998. *Multivariate Data Analysis*. 5th ed. USA, NJ: Prentice-Hall International.
- Hair, J., Black, W., Babin, B., and Anderson, R. 2010. *Multivariate Data Analysis*. 7th ed. USA, NJ: Prentice-Hall International.
- Hamilton, P. 2008. 'Frederick Burwick and James C. McKusick, eds, Faustus: From the German of Goethe, Translated by Samuel Taylor Coleridge. Oxford University Press'. *Angermion: A Yearbook of Anglo-German Cultural Relations* 1: 175-9.
- Haney, J. L. 1902. *The German Influence on S.T. Coleridge*. Philadelphia: The new era printing company.
- Hauhart, W. F. 1909. *The Reception of Goethe's Faust in England*. New York: The Columbia University Press.
- Hardcastle, R. A. 1993. 'Forensic Linguistics: an assessment of the CUSUM method for the Determination of authorship'. *Journal of the Forensic Science Society* 33: 95-106.
-1997. 'Cusum: A credible method for the determination of authorship'. *Science and Justice* 37: 129-138.
- Hastie Trevor, Tibshirani Robert, and Friedman Jerome. 2013. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer Business Media.
- Hatherall, D. and Hatherall, G. 1995. *Colloquial German*. Devon: Routledge, Taylor and Francis Group.
- Hausner, M. 1965. *A Vector Space Approach to Geometry*. USA: Dover publications Inc.

- Haykin, S. 1999. *Neural Networks. A Comprehensive Foundation*. London: Prentice Hall International.
- Hennig, C. 2007. 'Cluster-wise assessment for cluster stability'. *Computational Statistics and Data Analysis* 52: 258-71.
- Herdan, G. 1960. *Type- Token Mathematics*. The Hague: Mouton & Co.
1964. *Quantitative Linguistics*. London: Butterworth.
- Hilton, M. L. and Holmes, D. I. 1993. 'An assessment of cumulative sum charts for authorship attribution'. *Literary and Linguistic Computing* 8: 73-80.
- Hinneburg, A., Aggarwal, C., and Keim, D. 2000. 'What is the nearest neighbour in high dimensional spaces?' *ACMDL* 506: 515.
- Hirst, G., and Feiguina, O. 2007. 'Bigrams of syntactic labels for authorship discrimination of short texts'. *Literary and Linguistic Computing* 22: 405-417.
- Hockey, S. 2007. 'The history of humanities computing'. In Schreibman, S., Siemens, R., and Unsworth, J. (eds) *A Companion to Digital Humanities*. Oxford: Blackwell. 1-19.
- Holmes, D. I and Forsyth, R. 1995. 'The Federalist Revisited: New directions in Authorship Attribution'. *Literary and Linguistic Computing* 10: 111-127.
- Holmes D. I. and Tweedie F. J. 1995. 'Forensic Stylometry: A Review of the Cusum Controversy'. *Revue Informatique et Statistique dans les Sciences Humaines* 19-47.
- Holmes, D., Robertson, M., and Paez, R. 2001a. 'Stephen Crane and the New York Tribune: A case study in traditional and non-traditional authorship attribution'. *Computers in the Humanities* 35: 315-331.
- Holmes, D. I, Gordon, I. and Wilson, C. 2001b. 'A widow and her soldier: Stylometry and the American civil war'. *Literary and Linguistic Computing* 16: 403-420.

- Holmes, D. I. 1985. 'The analysis of literary style: a review'. *The Journal of the Royal Statistical Society* 148: 328-341.
-1991. 'Vocabulary richness and the prophetic voice'. *Literary and Linguistic Computing* 6: 259-268.
-1992. 'A stylometric analysis of Mormon scripture and related texts'. *Journal of the Royal Statistical Society*, 155: 91-120.
-1994. 'Authorship attribution'. *Computers and the Humanities* 28: 87-106.
-1998. 'The Evolution of Stylometry in humanities scholarship'. *Literary and Linguistic Computing* 13: 111-117.
-2003. 'Stylometry and the Civil War: the Case of the Pickett Letters'. *Chance* 16: 18-25.
- Hollmen Jaakk. 1996. 'Self- Organizing Map'. Last accessed 12 March 2014, from: <http://users.ics.aalto.fi/jhollmen/dippa/node9.html>.
- Honore, A. 1979. 'Some simple measures of richness of vocabulary'. *Association for Literary and Linguistic Computing* 7: 172-177.
- Hoorn, J., Frank, S., Kowalczyk, W., and van der Ham, F. (1999). 'Neural network identification of poets using letter sequences'. *Literary and Linguistic Computing* 14: 311-338.
- Hoover, L.D. 2001. 'Statistical stylistics and authorship attribution: An empirical investigation'. *Literary and Linguistic Computing* 16: 421-443.
-2002. 'Frequent Word Sequences and Statistical Stylistics'. *Literary and Linguistic Computing* 17: 157-180.
-2003a. 'Another perspective on vocabulary richness'. *Computer and the Humanities* 37: 151-178.
-2003b. 'Frequent Collocations and Authorial Style'. *Literary and Linguistic Computing* 18: 261-286.

-2003c. 'Multivariate Analysis and the Study of Style Variation'. *Literary and Linguistic Computing* 18: 341–360.
-2004. 'Testing Burrows' Delta'. *Literary and Linguistic Computing* 19: 453-475.
- Houvardas, J., and Stamatatos E. 2006. 'N-gram feature selection for authorship identification'. *Artificial Intelligence* 4183: 77-86.
- Hubert, Lawrence and Schultz, James. 1975. 'Hierarchical clustering and the concept of space distortion'. *British Journal of Mathematical and Statistical Psychology* 28: 121-133.
- Iqbal, F., Binsalleeh, B. C. Fung, and Debbabi, M. 2010. 'Mining Writerprint from Anonymous E-mails for Forensic Investigation'. *Digital Investigation* 7: 56-64.
- Izenman, A. 2008. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. New York: Springer-Verlag.
- Jain, A. K. 2003. 'Data Clustering: 50 Years beyond K means'. *Pattern Recognition Letters* 31: 651-66.
- Jain, A. and Dubes, R. 1988. *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice-Hall.
- Jain, A. and Moreau, J. 1987. 'Bootstrap technique in cluster analysis'. *Pattern Recognition* 20: 547-68.
- Jain, A. Murty, M., and Flynn, P. 1999. 'Data clustering: a review'. *ACM Computing Surveys* 31: 264-323.
- Jamak, A., Savatic, A., and Can, M. 2012. 'Principal component analysis for authorship attribution'. *Business Systems Research Journal* 3: 48-56.
- Jobson John. 1999. *Applied Multivariate Data Analysis: Regression and Experimental Design*. New York: Springer.

- Jockers Matthew and Witten Daniela. 2010. 'A Comparative Study of Machine Learning Methods for Authorship Attribution'. *Literary and Linguistic Computing* 25: 215–223
- Johnson, R. A. and Wichern, D.W. 1992. *Applied Multivariate Statistical Analysis*. 3rd ed. Englewood Cliffs, New Jersey: Prentice Hall.
- Jones, R. and Tschirner, E. 2006. *A Frequency Dictionary of German: Core Vocabulary for Learners*. New York: Routledge, Taylor and Francis Group
- Jolliffe, I. 2002. *Principal Components Analysis*. 2nd ed. Berlin: Springer.
- Jonthan Hope. 1994. *The Authorship of Shakespeare's Plays*. USA: Cambridge University Press.
- Juola, P. and Baayen, H. 2005. 'A controlled-corpus experiment in authorship attribution by cross-entropy'. *Literary and Linguistic Computing* 20: 59–67.
- Juola, P., Sofko, J. and Brennan, P. 2006. 'A prototype for authorship attribution studies'. *Literary and Linguistic Computing* 21: 169-178.
- Juola, P. 2004. 'Ad-hoc authorship attribution competition'. Last accessed 10 August 2009, from: http://www.mathcs.duq.edu/~juola/authorship_contest.html.
-2006. 'Authorship attribution'. *ACMDL1*: 233-334.
- Kaban, A. 2012. 'Distance Concentration and Detection of Meaningless Distances'. Last accessed 04 August 2013, from: <http://www.cs.bham.ac.uk/~axk/Dagstuhl.pdf>.
- Kachigan, S. 1991. *Multivariate Statistical Analysis: A conceptual introduction*. New York: Radius Press.
- Kantardzic Mehmed. 2011. *Data Mining: Concepts, Models, Methods, and Algorithms*. Piscataway, New Jersey: IEEE Press.
- Karlgren, J. and Cutting, D. 1994. 'Recognizing text genres with simple metrics using discriminant analysis'. *ACMDL* 2: 1071-1075.

- Kaski Sami. 1997a. 'Self-Organizing Maps'. Last accessed 12 March 2014, from: <http://users.ics.aalto.fi/sami/thesis/node18.html>.
-1997b. 'Data Exploration Using Self-Organizing Maps'. Last accessed 12 March 2014, from: <http://users.ics.aalto.fi/sami/thesis/>.
- Kaski S. J. and Kohonen, T. 2000. *Methods for Exploratory Cluster Analysis*. Finland: Physica-Verlag HD.
- Kaufman, L. and Rousseeuw, P.J. 1990. *Finding Groups in Data: An introduction to Cluster Analysis*. Hoboken, New Jersey: John Wiley & Sons, Inc.
- Kelih, E., Gordana, A., Grzybeck, P., and Ernst, S. 2005. 'Classification of author and/or genre. The impact of word length'. In Weihs, C. and Wolfgang, G. (eds.) *Classification. The Ubiquitous Challenge*. New York: Springer. 498-505.
- Keselj, V., Peng, F., Cercone, N., and Thomas, C. 2003. *N- Gram-based author profiles for authorship attribution*. Last accessed 10 March 2012, from: https://wiki.eecs.yorku.ca/course_archive/2014..15/W/6339/_media/10_1_1.1.87.754.pdf.
- Kessler, B., Nunberg, G. and Schütze, H. 1997. 'Automatic detection of text genre'. *ACMDL* 32-38.
- Kestemont Mike. 2014. 'Function Words in Authorship Attribution from Black Magic to Theory'. In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLfL)* 59-66.
- Kettenring, J. 2006. 'The practice of cluster analysis'. *Journal of Classification* 23: 3-30.
- Khmelev, D. V. and Tweedie, F. J. 2002. 'Using Markov chains for identification of writers'. *Literary and Linguistic Computing* 16: 299-307.
- Kjell, B. 1994. 'Authorship determination using letter pair frequencies with neural network classifiers'. *Literary and Linguistic Computing* 9: 119-124.

- Kjetsaa, G. 1978. 'The battle of the Quiet Don: Another pilot study'. *Computers and the Humanities* 11:341-346.
-1979. 'And Quiet Flows the Don through the computer'. *Literary and Linguistic Computing* 248-256.
- Kjetsaa G., Gustavsson, S., Beckman, B., and Gil, S. 1984. *The Authorship of 'The Quiet Don'*. USA: Solum Forlag.
- Kiviluoto, K. 1996. 'Topology preservation in self-organizing maps'. *IEEE Transactions on Neural Networks* 8: 294-99.
- Kohonen, T. 2001. *Self-Organizing Maps*. 3rd ed. Berlin: Springer.
- Kohlbecher, G. 2008. 'I wonder if Coleridge was likely to get the following things so wrong...'. Last accessed 20 July 2009, from: http://www.friendsofcoleridge.com/kohlbecher_Faustus.doc.pdf.
- Koppel, M., Argamon, S., and Shimoni, A.R. 2002. 'Automatically categorizing written texts by author gender'. *Literary and Linguistic Computing* 17: 401-412.
- Koppel, M., and Schler, J. 2003. 'Exploiting stylistic idiosyncrasies for authorship attribution'. *In Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis* 69-72.
-(2004). 'Authorship verification as a one-class classification problem'. Last accessed 20 July 2008, from: <http://www.machinelearning.org/proceedings/icml2004/papers/415.pdf>.
- Koppel, M., Mughaz, D., and Schler, J. 2004. 'Text categorization for authorship verification'. Last accessed 20 August 2008, from: <http://rutcor.rutgers.edu/~amai/aimath04/SpecialSessions/Koppel-aimath04.pdf>.
- Koppel M., Schler J. and Zigdon K. 2005. 'Determining an author's native language by mining a text for errors'. *ACMDL* 624-628.

- Koppel, M., Schler, J., Argamon, S., and Messeri, E. 2006. 'Authorship attribution with thousands of candidate authors'. Last accessed 10 July 2008, from: <http://www.csie.ntu.edu.tw/~r95038/Try/paper/paper%20WebIR/p659-koppel.pdf>.
- Koppel, M., Schler, J., and Bonchek-Dokow, E. 2007. 'Measuring differentiability: Unmasking pseudonymous authors'. *Journal of Machine Learning Research* 8: 1261-1276.
- Koppel, M., Schler, J., and Argamon, S. 2009. 'Computational methods in authorship attribution'. *Journal of American Society for Information and Technology* 60: 9-26.
-(2012). 'The fundamental problem of authorship attribution'. *English Studies* 39: 284-291.
- Köppen, M. 2000. 'The Curse of Dimensionality'. Last accessed 10 August 2014 from <http://www.yaroslavvb.com/papers/koppen-curse.pdf>.
- Korn, F., B. Pagel, and C. Faloutsos. 2001. 'On the Dimensionality Curse and the Self-Similarity Blessing'. *IEEE Transactions on Knowledge Data Engineering* 13: 96-111.
- Kruskal, J. B. 1964. 'Multidimensional Scaling by Optimizing Goodness of Fit to a nonmetric hypothesis'. *Psychometrika* 29: 1-27.
- Kruskal, J.B. and Wish, M. 1978. 'Multidimensional Scaling'. Last accessed 20 March 2014, from: <https://us.sagepub.com/en-us/nam/multidimensional-scaling/book432>.
- Krzanowski, W. J. 1988. *Principles of Multivariate Analysis*. Oxford: Clarendon Press.
- Kukushkina, O.V., Polikarpov, A. A., and Khmelev, D. V. 2002. 'Using literal and grammatical statistics for authorship attribution'. *Problems of Information Transmission* 37: 172-184.
- Kumar, T. S. (2004). 'Introduction to data mining'. Last accessed 02 March, 2010, from: http://www-users.cs.umn.edu/kumar/ch2_data.pdf.

- Ladyman, J. 2002. *Understanding Philosophy of Science*. London: Routledge.
- Lamber Maarten and Veenman Cor J. 2009. 'Forensic Authorship Attribution Using Compression Distances to Prototypes'. *Computational Forensics* 5717: 13-24.
- Lancashire, I. 1997. 'Empirically Determining Shakespeare's idiolect'. *Shakespeare Studies* 25: 171-85.
-1989. 'Paradigms of Authorship'. *Shakespeare Studies* 26: 298-301.
- Lance, G. N. and Williams, W. T. 1967. 'A general theory of classificatory sorting strategies I. hierarchical systems'. *ComputerJournal* 9: 373-80.
- Lange, T. et al. 2004. 'Stability-based validation of clustering solutions'. *Neural Computation* 16: 1299-1323.
- Laufer, B. and Goldstein, Z. 1995. 'Testing vocabulary knowledge: Size, strength, and computer adaptiveness'. *Language Learning* 54: 399-436.
- Lay, D. 2010. *Linear Algebra and its Applications*. 4th ed. London: Pearson.
- Lebart, L., Salem, A., and Berry, L. 1998. *Exploring Textual ata*. Netherlands: Kluwer Academic Publishers.
- Ledger, G. and Merriam, T. 1994. 'Shakespeare, Fletcher, and the two noble Kinsmen'. *Literary and Linguistic Computing* 9: 119-124.
- Ledger, G. 1995. 'An exploration of differences in the Pauline Epistles using multivariate statistical analysis'. *Literary and Linguistic Computing* 10: 85-97.
- Lee, J.A. and Verleysen, M. 2007. *Nonlinear Dimensionality Reduction*. New York: Springer science and business media.
- Lee, J. 2010. *Introduction to Topological Manifolds*. 2nd ed. Berlin: Springer.
- Lehmann, E. L. and Romano, J. P. 2005. *Testing Statistical Hypotheses*. 3rd ed. USA: Springer.
- Lehmann, E. L. 1997. 'Testing Statistical Hypotheses: The Story of a Book'. *Statistical Science* 12: 48-52.

- Levine, E. and Domany, E. 2001. 'Resampling method for unsupervised estimation of cluster validity'. *Neural Computation* 13: 2573-93.
- Lomenzo, A. J. 2008. 'One response to "Faust pas"'. Last accessed 10 March 2009, from: <http://blogs.law.harvard.edu/houghtonmodern/2008/05/20/faust-pas/>.
- Love Harold. 2002. *Authorship Attribution: An Introduction*. UK: Cambridge University Press.
- Lüdeling, A. and Kytö, M. 2009. *Corpus Linguistics: An International Handbook*. Berlin, Germany: Walter de Gruyter.
- Luyckx, K. 2010. 'Salability issues in authorship attribution'. Last accessed 20 March 2011, from: http://www.clips.uantwerpen.be/~kim/Papers/PhD-KimLuyckx_bookVersion.pdf.
-2004. 'Syntax-based features and machine learning techniques for authorship attribution'. Last accessed 20 March 2011, from: http://www.clips.ua.ac.be/stylometry/Papers/MAThesis_KimLuyckx.pdf.
- Luyckx, K. and Daelemans, W. 2005. 'Shallow text analysis and machine learning for authorship attribution'. Last accessed 20 March 2011, from: <http://www.clips.ua.ac.be/~kim/Papers/LD05.pdf>.
-2008. 'Authorship attribution and verification with many authors and limited data'. Last accessed 20 March 2011, from: <http://www.cnts.ua.ac.be/sites/default/files/ld08bnaic.pdf>.
-2011. 'The effect of author set and data size in authorship attribution'. *Literary and Linguistic Computing* 26: 35-55.
- Madigan, D., Genkin, A., Lewis, D., Genkin, Er, Argamon, S., and Ye, L. 2005. 'Author identification on the large scale'. Last accessed 10 June 2012, from: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.60.5324&rep=rep1&type=pdf>.

- Maguire, W. and McMahon, A. 2011. *Analysing Variation in English*. UK: Cambridge university press
- Mahlberg Michael. 2013. *Corpus Stylistics and Dickens's Fiction*. UK: Routledge, Taylor and Francis.
- Maindonald, J. and Braun, W. 2003. *Data analysis and Graphics Using R: An Example-Based Approach*. USA: Oxford University Press.
- Manning, C. and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. USA: MIT Press.
- Manning, C., Raghavan, P., and Schütze, H. 2008. *Introduction to Information Retrieval*. England: Cambridge University Press.
- Mannion David and Dixon Peter. 2004. 'Sentence- length and Authorship Attribution: The Case of Oliver Goldsmith'. *Literary and Linguistic Computing* 19: 497-508.
- Marcus, M. and Minc, H. 1988. *Introduction to Linear Algebra*. New York: Dover Publications.
- Mardia, K. V., Kent, J. T., and Bibby, J. M. 1979. *Multivariate Analysis*. London: Academic Press.
- Marcelo Luiz Brocardo, Issa Traore, Sherif Saad, Isaac Woungang. 2013. 'Authorship Verification for short Messages Using Stylometry'. Last accessed 10 June 2014, from: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6705711>.
- Marshall, D. 2009. 'Linear Algebra, Vectors, and Matrices'. Last accessed 21 May 2012, from: http://www.cs.cf.ac.uk/Dave/CM0167/Lectures/CM0167_Chap04_Linear_Algebra.pdf.
- Marshall H. Marshall. 1960. *Byron, Shelley, Hunt, and the Liberal*. New York: University of Pennsylvania Press.

- Martindale, C. and McKenzie, D. 1995. 'On the utility of content analysis in author attribution: The 'Federalist''. *Computers and the Humanities* 29: 259-270.
- Martinez, W., Martinez, A., and Solka, J. 2011. *Exploratory Data Analysis with Matlab*. London: CRC Press.
- Mascol, C. 1888a. 'Curves of Pauline and Pseudo-Pauline style I'. *Unitarian Review* 30: 452-460.
-1888b. 'Curves of Pauline and Pseudo-Pauline style II'. *Unitarian Review* 30: 539-546.
- Matthews, R., and Merriam, T. 1993. 'Neural computation in stylometry: An application to the works of Shakespeare and Fletcher'. *Literary and Linguistic Computing* 8: 203-209.
- Mazzeo Tilar J. 2006. *Plagiarism and Literary Property in the Romantic Period*. New York: University of Pennsylvania Press.
- Mays, J.C.C. (ed.) 2001. *The Collected Works of Samuel Taylor Coleridge Poetical Works*. USA: Princeton University Press.
-2007. 'Are Coleridge's plays worth the candle?'. *The Coleridge Bulletin N.S.* 29: 1-16.
-2012. 'Faustus on the table at Highgate'. *The Wordsworth Circle* 43: 119-127.
- McArthur, N. 2011. *Goethe's Faust Translated by Samuel Taylor Coleridge*. Manitoba: Strahan and Sons publishers.
- McEnery, T. and Wilson, A. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- McEnery, T. and Oakes M. 2000. 'Authorship identification and computational Stylometry'. In Dale, R., Moisl, H., and Somers, H. (eds) *Handbook of Natural Language Processing*. USA: Marcel Dekker, Inc. 234-248.

- McMenamin, G. R. 1993. *Forensic Stylistics*. Amsterdam: Elsevier.
- Mealand D. 1997. 'Measuring Genre differences in Mark with Correspondence Analysis'. *Literary and Linguistic Computing* 12: 227-245.
-1995. 'Correspondence Analysis of Luke'. *Literary and Linguistic Computing* 10: 85-98.
- Mehrotra K, Mohan, C. K., and Ranka, S. 1997. *Elements of Artificial Neural Networks*. Cambridge: MIT Press.
- Mehler, A., Pustynnikov, O., and Diewald, N. 2011. 'Geography of Social ontologies: Testing a variant of the Sapir-Whorff Hypothesis in the context of Wikipedia'. *Computer Speech and Language* 25: 716-40.
- Mendelson, B. 1975. *Introduction to Topology*. Boston: Allyn and Bacon.
- Mendenhall, T. C. 1887. 'The characteristic curves of composition'. *Science* 11: 237-249.
-1901. 'A mechanical solution to a literary problem'. *Popular Science* 9: 97-110.
- Merkel, D. 1997. 'Exploration of text collections with hierarchical feature maps'. *ACMDL* 31: 186-195.
- Merkel, D. and Rauber, A. 1997a. 'Alternative Ways for cluster visualization in self-organizing maps'. *In Proceedings of the Workshop on Self-Organizing Maps (WSOM-97)* 106-11.
-1997b. 'Cluster Connections: A Visualization technique to reveal cluster boundaries in self-organizing maps'. *In proceedings of the 9th Italian Workshop on Neural Nets* 22-24.
- Merriam, T. 1988. 'Was Hand B. in Sir Thomas Moor Heywood's Autograph?' *Notes and Queries* 233: 455-458.
-1994. 'Letter frequency as a discriminator of authorship'. *Notes and Queries* 239: 467-469.

-1996. 'Marlowe's hand in Edward III'. *Literary and Linguistic Computing* 8: 59-72.
-1998. 'Heterogeneous authorship in early Shakespeare and the problem of Henry V'. *Literary and Linguistic Computing* 13: 15-28.
- Meyer, W. W. 1979. *Personal Communication with W. W. Meyer from Meyer Engineers*. Washington: Kirkland.
- Mikros, G.K. and Argiri, E. K. 2007. 'Investigating topic influence in authorship attribution'. In *proceedings of the Workshop on Plagiarism Analysis, Authorship Identification, and Near-duplicate Detection* 1-17.
- Mikros, G. K. 2007. 'Authorship attribution using discriminant function analysis: Exploring literary stylistic variation in five Modern Greek novels'. In *proceedings of the 5th Trier Symposium on Quantitative Linguistics* 6-8.
- Mirkin, B. 2011. *Core Concepts in Data Analysis: Summarization, Correlation, and Visualization*. Berlin: Springer.
-2013. *Clustering. A Data Recovery Approach*. London: CRC Press.
- Moisl, H. and Jones, V. 2005. 'Cluster analysis of the Newcastle electronic corpus of Tyneside English: A comparison of methods'. *University of Twente* 20: 125-146.
- Moisl, H., Maguire, W., and Allen, W. 2006. 'Phonetic Variation in Tyneside: Exploratory Multivariate Analysis of the Newcastle Electronic Corpus of Tyneside English'. In Hinskens, F. (ed.) *Language Variation. European Perspectives*. Amsterdam: John Benjamins Publishing Co. 127-141.
- Moisl, H. 2008. 'Data Normalization for Variation in Document Length in Exploratory Multivariate Analysis of Text Corpora'. In *Proceedings of INFOS2008: 6th International Conference on Informatics and Systems, Cairo University* 27-29.

- Moisl, H. 2009a. 'Exploratory multivariate analysis'. In A. Lüdeling and M. Kytö (eds) *Corpus Linguistics: An International Handbook*. Berlin: de Gruyter Mouton. 874-899.
-2009b. 'Using electronic corpora to study language variation: the problem of data sparsity'. In Tsiplakou, S., Karyolemu, M., Pavlou, P. (eds) *Language Variation. European Perspectives*. Amsterdam: John Benjamins Publishing. 169-178.
-2009d. 'Sura length and lexical probability estimation in cluster analysis of the Qur'an'. *ACMDL* 8: 1-19.
-2009e. 'Using electronic corpora in historical dialectology research: the problem of document length variation'. In Dossena, M. and Lass, R. (eds) *Studies in English and European Historical Dialectology*. Berlin: Peter Lang. 67-90.
-2015. *Cluster Analysis for Corpus Linguistics*. Berlin: De Gruyter Mouton.
- Morton, A. Q. 1965. 'The authorship of Greek prose'. *Journal of the Royal Statistical* 128: 169-233.
-1978. *Literary Detection*. East Grinstead: Bowker Publishing.
-1986. 'A test of authorship based on words which are not repeated in the sample'. *Literary and Linguistic Computing* 1: 1-8.
- Morton, A. Q. and Michaelson, S. 1990. *The Qsum Plot*. University of Edinburgh.
- Mosteller, F. and Wallace, D. L. 1964. 'Inference and disputed authorship: The Federalist'. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute* 34: 277-279.
-1984. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. New York: Springer-Verlag Inc.
- Munkres, J. 2000. *Topology*. 2nd ed. London: Pearson Education International.

- Murray Christopher John. 2004. *Encyclopedia of the Romantic Era, 1760-1850*. New York, NJ: Fitzroy Dearborn.
-2009a. 'Give it up in despair: Coleridge and Goethe's Faust'. *Romanticism* 15: 1-15.
-2009b. 'Frederick Burwick and James C. McKusick, eds, Faustus: From the German of Goethe, Translated by Samuel Taylor Coleridge. Oxford University Press, 2007'. *BARS Bulletin and Review* 24, 66-67.
- Murriel St. Glare Byrne. 1932. 'Bibliographical Clues in Collaborative Plays'. *The Library* 13: 21-48.
- Nesterenko Alexander. 1979. *Values in Science: A Study in Operant Subjectivity*. University of Iowa.
- Nieto, V. 2004. 'Authorship attribution with the help of language engineering'. Last accessed 18 July 2008, from: <http://www.nada.kth.se/kurser/2D1418/uppsatser04/victor.pdf>.
- Oakes, M. P. 2014. *Literary Detective Work on the Computer*. Amsterdam: The Netherlands John Benjamins.
-2008. 'Corpus Linguistics and Stylometry'. Last accessed 12 June 2012, from: [http://pers-www.wlv.ac.uk/~in4326/papers/\\$U50.pdf](http://pers-www.wlv.ac.uk/~in4326/papers/$U50.pdf).
-1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- O'Donnell, B. 1966. 'Stephan Crane's The O'Ruddy: A problem in authorship discrimination'. In Leed, J. (ed.) *The Computer and Literary Style*. Kent State University Press. 107-115.
- Oja, E. and Kaski, S. 1999. *Kohonen Maps*. London: Elsevier.
- Oliphant, E. H. C. 1929. 'How Not to Play the Game of Parallels'. *Journal of English and Germanic Philology* 28: 1-15.

- Oliveira, M. and Levkowitz, H. 2003. 'From visual data exploration to visual data mining: A survey'. *IEEE Transactions on Visualization and Computer Graphics* 9: 378-394.
- Olsson, J. 2006. 'Using groups of common textual features for authorship attribution'. Last accessed 18 April 2012, from: <http://www.thetext.co.uk/authorship.doc/osslon>.
- O'Neill Michael and Howe Anthony. 2013. *The Oxford Handbook of Percy Bysshe Shelley*. Great Britain: Oxford University Press.
- Pang Kevin. 2003. 'Self-Organizing Maps'. Last accessed 15 March 2014, from: <http://www.cs.hmc.edu/~kpang/nn/som.html>.
- Pascual, D. Pla, F., and Sanchez, J. 2010. 'Cluster validation using information stability measures'. *Pattern Recognition Letters* 31: 454-61.
- Paulin, R., Clair, W., and Shaffer, E. 2008. 'A Gentleman of Literary Eminence'. Last accessed 10 March 2009, from: <http://sas-space.sas.ac.uk/4530/1/stc-faustus-review.pdf>.
- Peng, R. and Hengartner, N. 2002. 'Quantitative analysis of literary styles'. *The American Statistician* 56: 175-185.
- Peng, F., Shuurmans, D., Keselj, V., and Wang, S. 2003. 'Language independent authorship attribution using character level language models'. *ACMDL1*:267-274.
- Peng, F., Schuurmans, D., and Wang, S. 2004. 'Augmenting Naive Bayes Classifiers with Statistical Language Models'. *Journal of Information Retrieval* 7: 317-345.
- Pennebaker W. James. 2011. 'Your Use of Pronouns Reveals Your Personality'. Last accessed 17 May 2011, from: <https://hbr.org/2011/12/your-use-of-pronouns-reveals-your-personality>.
- Platt, J. R. 1964. 'Strong Inference Science'. *Science*, 164: 347-353.

- Plozlbauer, G., Rauber, A., Dittenbach, M. 2005a. 'A vector field visualization technique for self-organizing maps'. *Lecture Notes in Computer Science* 3518: 399-409.
-2005b. 'Advanced Visualization Techniques for Self-Organizing Maps with graph-based methods'. *Lecture Notes in Computer Science* 3497: 75-80.
- Popper, K. 2002. *The Logic of Scientific Discovery*. 7th ed. London and New York: Routledge Classics.
- Priddy, K. L. and Keller, P. E. 2005. *Artificial Neural Networks: An Introduction*. USA: Spie Press.
- Pyle, D. 1999. *Data Preparation for Data Mining*. San Francisco, CA: Morgan Kaufmann Publishers.
- Ramezani Reza, Navid Sheydaei, and Mohsen Kahani. 2013. 'Evaluating the Effects of Textual Features on Authorship Attribution Accuracy'. Last accessed 20 March 2014, from: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6682828>.
- Raddy, Y. T. 1970. 'Isaiah and the computer: A preliminary report'. *Computers and the Humanities* 5: 65-73.
- Reiman Donald H. 1977. *The Bodleian Shelley Manuscripts*. USA: Nora Crook Timothy Webb.
- Reiman Donald and Neil Fraistat (eds). 2002. *Shelley's Poetry and Prose*. USA: W. W. Norton & Company.
- Rencher, A. C. and Christensen, W. F. 2012. *Methods of Multivariate Analysis*. 3rd ed. Hoboken, New Jersey: John Wiley and Sons Inc.
- Riba, A. and Ginebra, J. 2004. 'Diversity of vocabulary and homogeneity of style in Tirant Lo Blanc'. Last accessed 10 June 2012, from: http://lexicometrica.univ-paris3.fr/jadt/jadt2004/pdf/JADT_091.pdf.

- Rietveld, T. and van Hout R. 1993. *Statistical Techniques for the Study of Language and Language Behavior*. Berlin: Mouton de Gruyter.
- Ritter, H., Martinetz, T., and Schulen, K. 1992. *Neural Computation and Self Organizing Maps*. London: Addison-Wesley.
- Robertson, R. 2008. 'Faustus. From the German of Goethe translated by S.T. Coleridge'. *Translation and Literature*: 17: 247-250.
- Robertson, S. E. and Sparck K. Jones. 1994. *Simple, Proven Approaches to Text Retrieval*. UK: University of Cambridge.
- Rohlf, F. 1974. 'Methods of comparing classifications'. *Annual Review of Ecology and Systematics* 5: 10-13.
- Romesburg, C. H. 1984. *Cluster Analysis for Researchers*. USA: Wadsworth Inc.
- Roper, M. Fields, P. J., and Schaalje, G. B. 2012. 'Stylometric analyses of the book of Mormon: A short history'. *Journal of the Book of Mormon and Other Restoration Scripture* 21: 28–45.
- Rudman, J. 1998. 'The state of authorship attribution studies: Some problems and solutions'. *Computers and the Humanities* 31: 351-365.
- Sakai, Y. and Hashimoto, S. 2011. 'Four-dimensional mathematical data visualization via embodied four-dimensional space display system'. *FORMA* 26: 11-18.
- Salton, G., Wong, A., and Yang, C. 1975. 'A vector space model for automatic indexing'. *ACMDL* 18: 613-20.
- Salton, G. and McGill, M. 1983. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.
- Sampley Arthur M. 1933. 'Verbal Tests for Peele's Plays'. *Studies in Philology* 30: 473–96.

- Sanderson, C. and Guenter, S. 2006. 'Short texts authorship attribution via sequence Kernels, Markov chains and author unmasking: An investigation'. *Association for Computational Linguistics* 482-91.
- Sanford, A. J., Aked, J. F., Moxey, L. M., and Mullin, J. 1996. 'A critical examination of assumptions underlying the cusum technique of forensic Linguistics'. *Forensic Linguistics* 1: 151-167.
- Saw, J.G., Yang, M.C.K., Mo, T.C. 1984. 'Chebyshev inequality with estimated mean and variance'. *The American Statistician* 38: 130–132.
- Schaalje, G. B., Roper, M., and Fields, P. J. 2012. 'Stylometric analyses of the book of Mormon: A short history'. *Journal of the Book of Mormon and Other Restoration Scripture* 21: 28-45.
- Schnorr, V. and Forst, G. 1995. *Mastering German Vocabulary: A Thematic Approach*. Hauppauge NY: Barron's Educational Series
- Schmid, S. 2008. 'Frederick Burwick and James C. McKusick, eds. Faustus. From the German of Goethe translated by S.T. Coleridge'. Last accessed 10 May 2009, from: http://www.friendsofcoleridge.com/Schmid_Faustus.pdf.
- Schreibman Susan, Siemens Ray, and Unsworth John. 2004. *A Companion to Digital Humanities*. USA: Black Well Publishing Ltd.
- Scott, D. and Thompson, J. 1983. 'Probability density estimation in higher dimensions'. Last accessed 10 January 2014, from: <http://www.stat.rice.edu/~scottdw/ftp/Tech.Reps/ifna83.ps>.
- Sebastiani, F. 2002. 'Machine learning in automated text categorization'. *ACM Computing Surveys* 34: 1-47.
- Seletsky, O., Huang, T., and Henderson-Frost, W. 2007. 'The Shakespeare authorship question'. Last accessed 20 September 2012, from: <http://www.cs.dartmouth.edu/datamining/final.pdf>.

- Sharma, A. K. 2005. *Biostatistics*. India: Discovery Publishing House.
- Shimek Cary. 2007. 'Lost Literature: UM scholar helps unveil great poet's secret work'.
 Last accessed 14 July 2012, from:
<http://leg.mt.gov/bills/2009/Minutes/House/Exhibits/agh61a03.pdf>.
- Sichel, H. S. 1975. 'On a distribution law for word frequencies'. *Journal of the American Statistical Association* 70: 542-547.
-1986. 'Word frequency distributions and type-token characteristics'.
Mathematical Scientist 11: 45-72.
- Siemens Ray and Schreibman Susan. 2013. *A Companion to Digital Literary Studies*. UK:
 A John Wiley & Sons Ltd.
- Sigurd, B., Eeg-Olofsson, M, and Weijer, J. 2004. 'Word length, sentence length and frequency-ZIPF revisited'. *Studia Linguistica* 58: 37-52.
- Silva Bruno. 2008. 'A Study of a Hybrid Parallel Self-Organizing Map Algorithm for large maps in Data Mining'. Last accessed 11 June 2013, from:
<http://ssdi.di.fct.unl.pt/~nmm/MyPapers/SM07.pdf>.
- Simpson, E. H. 1949. 'Measurement of diversity'. *Nature* 163-688.
- Simpson, J. 1979. *Matthew Arnold and Goethe*. England: W.S. Maney & Son Limited.
- Singh Jaswinder pal. 2012. 'Principal Components Analysis'. Last accessed 20 May 2013, from: <https://www.cs.princeton.edu/~jps/>.
- Singhal, A. Salton, G., and Buckley, C. 1995. 'Length normalization in degraded text collections'. Last accessed 17 July 2009, from: <http://singhal.info/ocr-norm.pdf>.
- Singhal, A., Buckley, C., and Mitra, M. 1996a. 'Pivoted document length normalization'. Last accessed 17 July 2009, from: <http://singhal.info/pivoted-dln.pdf>.
- Singhal, A., Salton, G., Mitra, M., and Buckley, C. 1996b. 'Document length normalization'. *Information Processing and Management* 32: 619-633.

- Singhal, A. 2001. 'Modern information retrieval: A brief overview'. *IEEE Data Engineering Bulletin* 24: 35-43.
- Smiles, S. LL.D. 1891. *Memoir and Correspondence of the Late John Murray with an Account of the Origin and Progress of the House, 1768-1843. In Two Volumes- Vol. I with portraits*. London: John Murray.
- Smith, M.W.A. 1983. 'Recent experience and new developments of methods for the determination of authorship'. *Association for Literary and Linguistic Computing Bulletin* 11: 73-82.
-1985. 'An investigation of the basis of Morton's method for the determination of authorship'. *Style* 19: 341-368.
- Smith, R. and Rochester, N. Y. 2012. 'Distinct word length frequencies: distributions and symbol entropies'. *Glottometrics* 23: 7-22.
- Smith Jordan. 2008. 'Issues in Text Similarity and Categorization'. Last accessed 10 March 2014, from: http://www.music.mcgill.ca/jordan/coursework/mumt611/text_categorization.ppt.
- Smith Peter W. H. and Jong, Gea de. 2005. 'Speaker Identification: Function Words and Beyond'. Last accessed 10 March 2014, from: http://www.slidefinder.net/s/speaker_identification_function_words_beyond/iafl2005/9126753.
- Sneath, P. and Sokal, R. 1963. *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. London: Freeman.
- Soboroff, I., Nicholas, C., Kukla, J., and Ebert, D. 1998. 'Visualizing document authorship using N-grams and latent semantic indexing'. *ACMDL* 43-48.
- Soffer Walter. 1987. *From Science to Subjectivity. An Introduction of Descartes' Meditations*. USA: Praeger.

- Sokal, R. and Rohlf, F. 1962. 'The comparison of dendrograms by objective methods'. *Taxon* 11: 33-40.
- Somers, H. 1998. 'An attempt to use weighted cusums to identify sublanguages'. *New Methods in Language Processing and Computational Natural Language Learning* 131-139.
- Somers, H. and Tweedie, F. 2003. 'Authorship attribution and pastiche'. *Computers and the Humanities* 37: 407-429.
- Stamatatos, E., Kokkinakis, G., and Fakotakis, N 1999. 'Automatic authorship attribution'. Last accessed 12 August 2010, from: <https://www.aclweb.org/anthology/E/E99/E99-1021.pdf>.
-2000. 'Automatic text categorization in terms of genre and author'. *Association for computational linguistics* 26: 471-495.
-2001. 'Computer-based authorship attribution without lexical measures'. *Computer and the humanities* 35: 193- 214.
- Stamatatos, E. 2008. 'Author identification: Using text sampling to handle the class imbalance problem'. *Information Processing and Management* 44: 790-799.
-2009. 'A survey of modern authorship attribution methods'. *Journal of the American society for information science and technology* 60: 538-556.
-2013. 'On the Robustness of Authorship Attribution based on Character N-Gram Features'. *Journal of Law and Policy* 22: 421-439.
- Stańczyk, U. and Cyran A. 2007. 'Machine Learning Approach to Authorship Attribution of Literary Texts'. *International journal of applied mathematics and informatics* 4: 151-158.
- Steen, V. K. 2012. *Elements of Statistics*. Belgium: University of Liege.

- Steinbach, M., Ertöz, L. and Kumar, V. 2004. 'The Challenges of Clustering high-dimensional data'. In Wille, L. (ed.) *New Directions in Statistical Physics*. Berlin: Springer. 273-309.
- Stillinger, J. 1994. *Coleridge and Textual Instability*. USA: Oxford University Press.
- Stockburger, D. W. 1997. *Multivariate Statistics: Concepts, Models, and Applications*. USA: David. W. Stockburger.
- Stokoe, F. W. 1926. *German Influence in the English Romantic Period, 1788-1818: with Reference to Scott, Coleridge, Shelley, and Byron*. New York: University Press of Cambridge.
- Tabachnik, B. and Fidell, L. 2001. *Using Multivariate Statistics*. 4th ed. Boston, MA: Allyn and Bacon.
- Tallentire, D.R. 1972. *An Appraisal of Methods and Models in Computational Stylistics, with Particular Reference to Authorship Attribution*. UK: University of Cambridge.
- Tanguy, L., Urieli, A., Calderone, B., Hathout, N., and Sajous, F. 2011. 'A multitude of linguistically-rich features for authorship attribution'. Last accessed 11 March 2011, from: <http://ceur-ws.org/Vol-1177/CLEF2011wn-PAN-TanguyEt2011.pdf>.
- Tenenbaum, J. deSilva, V., and Langford, J. 2000. 'A Global geometric framework for nonlinear dimensionality reduction'. *Science* 290: 2319-23.
- Thoiron, P. 1986. 'Diversity Index and Entropy as Measures of Lexical Richness'. *Computers and the Humanities* 20: 197-202.
- Timm Neil, H. 2002. *Applied Multivariate Analysis*. Berlin, Heidelberg: Springer-Verlag.
- Tweedie, F. J., Singh, S. and Holmes, D. I. 1996. 'Neural network applications in stylometry: The Federalist Papers'. *Computers and the Humanities* 30: 1-10.
- Tweedie, F., and Baayen, R. 1998. 'How variable may a constant be? Measures of lexical richness in perspective'. *Computers and the Humanities* 32: 323-352.

- Tweedie, F., Holmes, D. I., and Corns, T. 1998. 'The provenance of De Doctrina Christiana, attributed to John Milton: A statistical investigation'. *Literary and Linguistic Computing* 13: 77-87.
- Thomas, L., Pfister, H., and Peterson, P. 2004. 'Issues related to the construction of a purpose-built domain specific word corpus'. *Australian Journal of Educational and Developmental Psychology* 4: 13-28.
- Thorndike, A. H. 1901. *The influence of Beaumont and Fletcher on Shakespeare*. New York: AMS Press
- Tinsley, H. and Brown, S. 2000. *Handbook of Applied Multivariate Statistics and Mathematical Modelling*. New York: Academic Press.
- Turley R. Marggraf. 2009. *Bright Stars: John Keats, Bary Cornwall and Romantic Literary Culture*. UK: Liverpool University Press.
- Taylor, B. 1890. 'Faust: A Tragedy Translated Into English, In The Original Metres'. Last accessed 28 August 2015, from: <https://www.gutenberg.org/files/14591/14591-h/14591-h.htm>.
- Uhlig. S. H. 2010. 'Faustus from the German of Goethe translated by S.T. Coleridge'. *The Review of English Studies* 61: 645-648.
- Ultsch, A. and Siemon, H. 1990. 'Kohonen's self-organizing feature maps for exploratory data analysis'. Last accessed 12 March 2014, from: <https://www.uni-marburg.de/fb12/datenbionik/pdf/pubs/1990/UltschSiemon90>.
- Van Halteren, H. 2004. 'Linguistic profiling for authorship recognition and verification'. *ACMDL* 199: 1-8.
-2007. 'Author verification by linguistic profiling: An exploration of the parameter space'. *ACM Transactions on Speech and Language Processing* 4: 1-17.

- Van Halteren, H., Baayen, H., Tweedie, F., Haverkort, M. and Neijt, A. 2005. 'New machine learning methods demonstrate the existence of a human stylometric'. *Journal of Quantitative Linguistics* 12: 65-77.
- van Rijsbergen, C. 1979. *Information Retrieval*. 2nd ed. London: Butterworths.
- van Tongeren, O. F. R. 1995. 'Cluster Analysis'. In R. H. G. Jongman, C. J. F. Ter Braak and O. F. R. van Tongeren (eds) *Data Analysis in Community and Landscape Ecology*. New York: Cambridge University Press. 174-212.
- Venna, J. and Kaski, S. 2001. 'Neighborhood preservation in nonlinear projection methods: an experimental study'. *Lecture Notes in Computer Science* 2130: 458-91.
- Verleysen, M., François, D., Simon, G., and Wertz, V. 2003. 'On the effects of dimensionality on data analysis with neural networks'. In J. Mira (ed.) *International Work-Conference on Artificial and Natural Neural Networks*. Mao, Menorca (Spain). 105-112.
- Verleysen, M. and Francois, G. 2005. 'The curse of dimensionality in data mining and time series prediction'. *Lecture Notes in Computer Science* 3512: 758-770.
- Verleysen, M. 2003. 'Learning high-dimensional data'. Last accessed 16 May 2010, from: <http://perso.uclouvain.be/michel.verleysen/papers/nato03mv.pdf>.
- Vesanto, J. 1999. 'SOM based data visualization methods'. *Intelligent Data Analysis* 3: 111-26.
-2000. 'Using SOM in Data Mining'. Last accessed 12 May 2013, from: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.33.942>.
- Vickers Brian. 2002. *Counterfeiting's Shakespeare: Evidence, Authorship, and John Ford's Funeral Elegy*. UK: Cambridge University Press.
- Volk, M. 2006. 'German Prepositions and Their Kin'. In Saint-Dizier, P. (ed.) *Syntax and Semantics of Prepositions*. The Netherlands: Springer. 83-99.

- Wang, J. T. L., Zaki, L., Toivonen, H. T. T., and Shasha, D. 2005. *Data Mining in Bioinformatics*. USA: Springer.
- Wang, L. and Shen, X. 2006. 'Multi-category Support Vector Machines, Feature Selection, and Solution Path'. *Statistica Sinica* 16: 617-633.
- Ward, J.H. 1963. 'Hierarchical grouping to optimize an objective function'. *Journal of the American Statistical Associations* 58: 236-244.
- Waugh, S., Adams, A., and Tweedie, F. 2000. 'Computational stylistics using artificial neural networks'. *Literary and Linguistic Computing* 15: 187-198.
- Webb, A. 2002. *Statistical Pattern Recognition*. 2nd ed. New Jersey: John Wiley and Sons.
- Wegener, S. 2008. 'Literature fingerprinting: A new method for visual literary analysis'. Last accessed 22 June 2010, from: <http://www.rw.cdl.uni-saarland.de/teaching/va08/session1/literaturefingerprints.pdf>.
- Wiechmann, D. 2008. *Cluster Analysis*. New York: John Wiley and Sons, Inc.
- Williams, C. B. 1970. *Style and Vocabulary*. New York: Hafner Publishing Co.
- Williamson Graham. 2009. 'Type-token ratio'. Last accessed 20 March 2013, from: <http://www.speech.therapyinformation>.
- Wing, W. and Chan, Y. 2006. 'A survey on multivariate data visualization'. Last accessed 10 August 2011, from: <http://www.cse.ust.hk/~wallacem/winchan/research/multivis-report-winnie.pdf>.
- Wishart, D. 1969. *FORTRAN II Programs for 8 Methods of Cluster Analysis (CLUSTAN I)*. The University of Kansas: Lawrence, State Geological Survey.
- Woods, A., Fletcher, P., and Hughes, A. 1986. *Statistics in Language Studies*. Cambridge: Cambridge University Press.
- Xu, J. and Croft, W. 2009. *Clustering*. Hoboken NJ: Wiley.

- Yang, Y. 1999. 'An evaluation of statistical approaches to text categorization'. Last accessed 11 June 2012, from: <http://www.csie.ntu.edu.tw/~b91082/irj99.pdf>.
- Yankov, D. and Keogh, E. 2006. 'Manifold Clustering of Shapes'. Last accessed 10 December 2014, from: <http://www.cs.ucr.edu/~eamonn/ManifoldShapeClustering.pdf>.
- Yu, B. 2012. 'Function Words for Chinese Authorship Attribution: A review of authorship Attribution'. *Workshop on Computational Linguistics for Literature* 45–53.
- Yu, L. and Han, Y. 2010. 'A variance reduction framework for stable feature selection'. *The ASA Data Science Journal* 5: 428-445.
- Yule, G. 1938. 'On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authorship'. *Biometrika* 30: 363-390.
1944. *The Statistical Study of Literary Vocabulary*. Cambridge, UK: Cambridge University Press.
- Zeng, X., Tan, M., and Chen, W. 2011. *Extraction of Visual Material and Spatial Information from Text Description from Scene Visualization*. Berlin, Heidelberg: Springer-Verlag.
- Zhang, H. H., Liv, Y., Wn, Y, and Zhu, J. 2008. 'Variable Selection for Mutli-category SVM via supnorm realization'. *Electronic Journal of Statistics* 2: 149-167.
- Zhao, Y. and Zobel, J. 2005. 'Effective and scalable authorship attribution using function words'. *Lecture Notes in Computer Science* 3689: 174-189.
- Zhao, Y. and Zobel, J. 2006. 'Sarching with style: Authorship attribution in classic literature'. Last accessed 11 August 2013, from: <http://goanna.cs.rmit.edu.au/~jz/fulltext/acsc07yz.pdf>.
- Zhao, Y., Zobel, J. and Vines, P. 2006. 'Using relative entropy for authorship attribution'. *Lecture Notes in Computer Science* 4182: 92-105.

- Zhao, Y. and CompSci, B. 2007. *Effective Authorship Attribution in Large Document Collections*. Australia: RMIT University.
- Zheng, R., Li, J., Chen, H. and Huang, Z. 2006. 'A framework for authorship identification of online messages: Writing-style features and classification techniques'. *Journal of the American Society for Information Science and Technology* 57: 378–393.
- Zhenshi Li. 2013. 'An Exploratory Study on Authorship Verification Models for Forensic Purposes'. Last accessed 29 March 2014, from: http://www.tbm.tudelft.nl/fileadmin/Faculteit/TBM/Over_de_Faculteit/Afdelingen/Afdeling_Infrastructure_Systems_and_Services/Sectie_Informatie_en_Communicatie_Technologie/medewerkers/jan_van_den_berg/news/doc/Master-Thesis-Final-Zhenshi-Li.pdf.
- Zipf, G. K. 1949. *Human Behavior and the Principle of Least Effort*. Cambridge, Mass.: Addison-Wesley Press.
-1965. *The Psychobiology of Language. An Introduction to Dynamic Philology*. Cambridge: MIT Press.