

HIDE AND SEEK: QUANTITATIVE AUTHORSHIP IDENTIFICATION IN LANGUAGE CONCEALMENT

by

HELGA WENDELBERGER

(Under the Direction of William A. Kretzschmar, Jr.)

ABSTRACT

This quantitative statistical investigation analyzes two different authorship case studies that entail language concealment. The investigation focuses on whether individuals can disguise their individual idiolects by changing the characteristics of their writing style, deliberately or unintentionally, or whether an individual's voice is so habitually ingrained with style markers that it can be identified even if disguised. The first study investigates unintentional language concealment involving psychic channeling claims. The second study determines what happens to style markers in a pseudonymic environment where authorship is deliberately disguised. In addition, this case study investigates how an individual's idiolect will react in a cross-genre situation (fiction versus non-fiction). The investigation stems from the hypothesis that an individual's idiolect contains characteristic markers, which can be used in scientific linguistic quantitative investigations to determine the authorship of written documents. Each multivariate technique that is used (principal component analysis, discriminant analysis, and cluster analysis) supplies a piece of the puzzle that ultimately completes the picture of the analysis. The compilation of the three techniques (referred to as the multivariate triad), used in each of the studies, establishes that the quantitative methodology used in conjunction with high frequency

function words is robust enough to predict authorship to a high probability provided that the sample size of the texts to be investigated is sizeable.

INDEX WORDS: Corpus Linguistics, Stylometry, Forensic Linguistics, Author Identification, Author Attribution, Multivariate Analysis, Genre, Parapsychology,

HIDE AND SEEK: QUANTITATIVE AUTHORSHIP IDENTIFICATION IN LANGUAGE
CONCEALMENT

by

HELGA WENDELBERGER

B.A., University of Georgia, 2002

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial
Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2006

© 2006

Helga Wendelberger

All Rights Reserved

HIDE AND SEEK: QUANTITATIVE AUTHORSHIP IDENTIFICATION IN LANGUAGE
CONCEALMENT

by

HELGA WENDELBERGER

Major Professor: William A. Kretzschmar Jr.

Committee: Marlyse Baptista
Renate Born

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2006

DEDICATION

To my father, George Wendelberger and my Oma, Therese Ghezin.

Two extraordinary souls that I lost too early in life but whose courage, strength and influence
remain the guide and moral compass of my being.

ACKNOWLEDGEMENTS

There are several people who I would like to acknowledge for their support during my incredible journey to obtain my doctoral degree.

I am deeply indebted to Dr. William Kretzschmar, who first introduced me to the concepts of author identification and corpus linguistics. His invaluable guidance and assistance made this dissertation possible. I would like to thank my committee members, Dr. Marlyse Baptista and Dr. Renate Born for allowing me to benefit from the wealth of their linguistic knowledge as well as agreeing to be on my committee. All three professors embody the best of what is academia and I shall always be grateful for the encouragement, compassion and respect they afforded me during this process.

I would like to thank Lynnette Lang and Debbie Vaughn, administrators for the Linguistic Program and the Linguistic Atlas office. Their knowledge of the inner-workings of the linguistic program is unsurpassed, but more importantly their invariable positive attitudes, and patience made any and all of my perceived administrative hurdles melt into oblivion. In addition, I would like to thank Dr. Christopher Hayes and Dr. Jared Klein for the graduate teaching assistantships that not only sustained me financially but also permitted me to do what I love – teach.

On the personal side, I would like thank my brother, Werner, my sister-in-law, Carol and my sister Ingrid. Werner's unwavering, non-judgmental support is unparalleled. He is my hero. I greatly appreciate Carol's consummate faith in my endeavor and Ingrid for her endearing belief that all dreams are possible.

I would like to acknowledge and thank some dear and cherished friends: Sue Carroll, Judy Amonette, Felicia Folds, Christine Creed-Kudelka, Linda Enis, Sandra Harding, and Marlene Kemp-Dynin. Without this incredibly precious nucleus of friends, I could not have embarked on this journey let alone completed it. I am truly grateful for each of their friendships. They are treasured.

In addition, I would like to thank some special friends from my previous life in corporate America: Lisa Moretti-Chakford, Michelle Robicheaux, Sarah Hanley, and Robert Connors. When I made the decision to return to school, many thought it was a passing whim or perhaps a unique expression of a mid-life crisis. Although these special friends might have privately questioned my decision, they trusted in my journey and expressed nothing less than their full support. I am thankful for their confidence.

Finally, I would like to give thanks for my remarkable dogs, Triste and Onyx. These two sweet spirits sat devotedly by my side, under my feet and sometimes on my feet for every word in this dissertation. Their unconditional love continues to be my inspiration.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	v
LIST OF TABLES	x
LIST OF FIGURES.....	xii
CHAPTER	
1 INTRODUCTION.....	1
Purpose of the Study.....	2
Study I – Psychic Language	3
Study II - Genre	4
Dissertation Organization	5
2 LITERATURE REVIEW.....	8
Author Identification - Quantitative	8
History	8
Function Words	13
Multivariate Analysis.....	18
Author Identification – Forensic Linguistics.....	21
Terminology: Register, Genre, Text Type.....	27
Genre.....	29
Genre – Author Identification.....	30
Genre – Multidimensional Studies	32

3	METHODOLOGY	36
	Corpora	37
	Optical Character Recognition	37
	WordSmith Tools.....	40
	Excel	41
	SPSS 13.0	42
	Multivariate Statistical Analysis.....	42
	Principle Component Analysis	43
	Discriminant Analysis	44
	Cluster Analysis.....	45
4	PSYCHIC LANGUAGE STUDY	47
	Problem Statement.....	47
	Background of the Study	48
	Parapsychology Expenditures.....	49
	Biography of Psychic Authors.....	51
	Parapsychology and Language	53
	Corpora of writings by Psychics.....	54
	Determining Word Frequency Hierarchy - Psychic	56
	Analysis Matrix – Psychic data	62
	Analysis – Psychic Data	65
5	GENRE STUDY	82
	Problem Statement.....	83
	Background of the Study	85

Pseudonymic Authorship.....	85
Biography of Genre Authors	86
Corpora - Genre	87
Determining Word Frequency Hierarchy - Genre	89
Analysis Matrix – Genre data	92
Analysis – Genre Data	93
6 CONCLUSION	109
REFERENCES.....	116
APPENDICES.....	134
A Psychic Matrix – Raw Data.....	134
B Genre Matrix – Raw Data	146

LIST OF TABLES

	Page
Table 3.1 Excel author segmentation sample – raw data	41
Table 3.2 Excel author segmentation sample – relative frequencies	41
Table 4.1 Hierarchy statistics – Psychic data	58
Table 4.2 Word and sentence length statistics	59
Table 4.3 Hierarchy word frequency list – Psychic data	60
Table 4.4 Sample concordance from WordSmith Tools – Psychic data	61
Table 4.5 Sample matrix raw data – Psychic data	63
Table 4.6 Word count discrepancies – Psychic data	64
Table 4.7 Sample matrix relative frequencies – Psychic data	64
Table 4.8 Total variance explained, principal components 1-2 – Psychic data	66
Table 4.9 Rotated component matrix – Psychic data	70
Table 4.10 Principal components parts of speech – Psychic data	71
Table 4.11 Total variance explained, principal components 1-13 – Psychic data	73
Table 4.12 Tests of equality group of group means – Psychic data	74
Table 4.13 Wilks' Lambda – Psychic data	76
Table 4.14 Classification results – Psychic data	76
Table 5.1 Expansion data of common contracted forms – Genre data	89
Table 5.2 Hierarchy statistic – Genre data	90
Table 5.3 Hierarchy word frequency list – Genre data	91

Table 5.4	Sample matrix raw data – Genre data	92
Table 5.5	Sample matrix relative frequencies – Genre data	93
Table 5.6	Total variance explained, principal components 1-2 – Genre data	94
Table 5.7	Rotated component matrix – Genre data.....	98
Table 5.8	Principal components 1 and 2 Parts of Speech – Genre data	99
Table 5.9	Total variance explained, principal components 1-10 – Genre data	101
Table 5.10	Tests of equality of group means – Genre data.....	102
Table 5.11	Wilks’ Lambda – Genre data	103
Table 5.12	Classification results – Genre data.....	104

LIST OF FIGURES

	Page
Figure 2.1a One-dimensional plot – four genres.....	34
Figure 2.1b Two-dimensional plot – four genres.....	34
Figure 4.1a Psychic data word plot.....	67
Figure 4.1b Psychic author segment behavior	68
Figure 4.1c Psychic discriminant analysis	77
Figure 4.1d Psychic dendrogram – cluster analysis.....	80
Figure 5.1a Genre data word plot.....	95
Figure 5.1b Genre author segment behavior	96
Figure 5.1c Genre discriminant analysis.....	105
Figure 5.1d Genre dendrogram – cluster analysis.....	107

CHAPTER 1

INTRODUCTION

The ubiquitous term “forensic” permeates our daily lives. Whether in media print, televised crime dramas, or in actual courtrooms, it has become the mantra of any investigative environment. What does the “forensic” evidence show? The 21st century is rife with “forensic” psychologists, “forensic” anthropologists, “forensic” entomologists, and other “forensic” scientists. Linguistics, the scientific study of language, is included in this “forensic” phenomenon. Forensic Linguistics has begun to flourish for the potential evidentiary analyses it can provide. Under this rubric, the scientific application of linguistics in a forensic context includes trademark violations, language crimes such as threats and bribery, and perhaps the best known, author identification.

Author identification using quantitative methods has a long and rich history in literary studies. Within the past two decades, however, author attribution has expanded to include linguistic analyses, which attempt to discern who may have written a text pertaining to legal issues involving threat assessment, language crimes, and slander - in other words, the “whodunit” scenario. This expansion has led some to claim that an individual’s language contains an identifiable fingerprint or that language is literally literary DNA.

Language variation studies confirm that no two members of a speech community use language in precisely the same way. This allows me to hypothesize that an individual’s idiolect contains habitual characteristic markers, which can be used in scientific linguistic quantitative

investigations to determine the authorship of various written documents. With the advent of computer technology, which allows linguists to build corpora, author identification using quantitative methods may determine with reasonable accuracy the probability of authorship.

Purpose of the study

This study analyzes two different authorship studies using specially built corpora and quantitative multivariate statistical methodology. Each study is designed to investigate specific questions about authorship identification. The purpose of this study is to see whether individuals can disguise their individual idiolects by changing the characteristics of their writing style, deliberately or unintentionally, or whether an individual's voice is so habitually ingrained with style markers that it can be identified even if disguised. As our field moves from literary studies to the forensic environment it is important that we understand how much control individuals have over their voices under different conditions.

Many author identification studies involve speculative suspect lists. For example, in the unresolved 1996 murder of six-year-old JonBenét Ramsey, suspicion continues to revolve around her parents, John and Patsy Ramsey. Literary studies continue to speculate whether someone other than Shakespeare wrote Shakespeare's works. Many studies have suggested that either Christopher Marlowe – playwright, or Sir Francis Bacon – philosopher/writer, might be the actual author.

Some analyses have an identified small nucleus of suspects, such as the investigation of Hamilton and Madison in the twelve disputed *Federalist Papers* where both claimed authorship. The *Federalist Papers* (77 essays) were published anonymously in 1787-88 by Alexander Hamilton, John Jay, and James Madison to persuade the citizens of New York to ratify the

constitution. Hamilton is the author of forty-three papers, Jay of five, and Madison of fourteen. The twelve disputed essays had two confirmed suspects: Hamilton and Madison. F. Mosteller set out to determine who the actual author was, Hamilton or Madison, using statistical quantitative methodology.

Other investigations focus on a retrospective investigation, where the case is resolutely solved, however, questions prevail as to what a linguistic examination could have provided the investigation had comparative documents been available. For example, the Unabomber case, from 1978 to 1996, Ted Kaczynski using the acronym “FC” terrorized college professors and executives with mail bombs. Kaczynski’s sister-in-law, Linda Patrik, recognized similarities between Ted’s privately written rantings and those of “FC” in a 35,000-word manifesto, published in 1995 by the *New York Times*, which ultimately led to his arrest and conviction.

Each type of study presents its own challenges. In a high profile case there is a chance a researcher might be unduly influenced from outside sources as to potential authorship. A limited data source might stymie potential investigations or limit them in their scope. Retrospection can demonstrate and highlight past issues or future concerns, and help us determine replicable successful author identification methods; however, it cannot predict if probable identification would have occurred at the time of the event.

Study I – Psychic Language

In designing my first study, candid consideration was given to these issues. I wanted a scenario involving author identification that had not been previously attempted. It was important that the set of circumstances entailed in my study carried no issues to which I had a prior bias. In addition, I wanted to do a case study where my findings would have practical value and

contribute to the field of author attribution. Finally, the study and the results could be applied and/or replicated in future authorship studies. The following scenario, which investigates author identification, allows me to achieve that. Parapsychology is currently a billion-dollar industry and growing exponentially each year as it gains more popularity. The areas found under the rubric of Parapsychology operate on the premise of personal beliefs and values that can rarely be scientifically investigated and corroborated. A quantitative study of the language involved in the parapsychology sub-field, psychic channeling, which claim spiritual entities that use language, provides an opportunity for a scientific investigation that focuses on author attribution as well as determines the validity of such claims. In 1971, Ruth Montgomery, a renowned psychic, claimed that her deceased friend, sensitive Arthur Ford, was continuing to speak through her in the form of automatic text writings and dictated an entire book, *A World Beyond*. Ruth Montgomery has said that she firmly believed the voice in *A World Beyond* was not her own. During his lifetime, Arthur Ford wrote several books, as did Ruth Montgomery. These texts, containing their individual language, supply us with the materials needed for a comparative author attribution investigation.

Question: Can a quantitative linguistic analysis determine whose voice is found in *A World Beyond* and substantiate psychic channeling claims?

Study II - Genre

My second study focuses not on author identification, but on author intention. I wanted to investigate if authors can deliberately disguise their individual idiolect by changing the characteristics of their writing style, principally in pseudonymic circumstances when the author traverses established genre lines. The following scenario provides the perfect forum upon which

to analyze how language will react in cross-genre situations and what we can expect as linguists under these circumstances.

In 1996 Joe Klein, a well-known political columnist and television political commentator, wrote the political novel *Primary Colors* under the pseudonym Anonymous to disguise his identity in a highly charged political environment. His deliberate attempt to conceal his identity as the writer of this unflattering roman à clef about President Bill Clinton's 1992 presidential campaign, consumed the world of United States politics for seven months.

- Question:**
1. What happens to the individual's voice in a pseudonymic environment and does it retain its habitual markers?
 2. Does the fact that Joe Klein's writings cross genres, from political essay to political novel, make it more difficult to ascertain his authorship?

Multivariate statistical methodology, comprised of principal component analysis, discriminant analysis and cluster analysis, will be used to determine authorship in Study I – Psychic language and to determine whether Joe Klein can retain his habitual markers in a pseudonymic environment that crosses genre in Study II - Genre.

Dissertation Organization

This dissertation is organized into five chapters: 2. Literature Review, 3. Methodology, 4. Psychic Language Study, 5. Genre Study, and 6. Conclusion. In chapter two, I review and discuss the various topics that entail an author attribution study using quantitative methodology. I describe the history of quantitative author identification analyses and provide an overview of the various quantitative methodologies available: successful and unsuccessful. In this chapter, I discuss the benefits of using function words in this type of study and I review the success of combining them with multivariate analyses. In addition, I address author identification in a forensic environment and the various methodologies that have been applied. Finally, I focus on

genre, its terminology, and how it affects author identification overall and specifically with multidimensional studies.

In chapter three, I focus on the various methodologies that I used to compile the data and complete the Psychic language study and the Genre study. I briefly discuss the corpora needed for each study. I review Optical Character Recognition (OCR), the process used to convert texts into electronic data that can be analyzed. In addition, in chapter three I provide an overview of WordSmith Tools, a software package that allows an analyst to see how words behave in a text, as well as briefly discuss Excel, the spreadsheet program used to develop the matrix in both studies and SPSS a software package that I used to perform statistical analyses on the psychic language data and the genre data. Finally, I address multivariate statistical analysis and the three specific statistical procedures used in this dissertation: principal component analysis, discriminant analysis, and cluster analysis.

Chapters four and five, present each individual study in its entirety. In each chapter, I outline how I built the individual study from its inception: choosing the texts, converting them into electronic analyzable data, building the corpora, and determining the hierarchy matrices needed for analysis. I provide a systematic method that can be replicated and/or employed in future studies. I then focus on the analysis and results for the three statistical procedures: principal component analysis, discriminant analysis and cluster analysis, which I refer to as the multivariate triad.

In chapter six, I discuss the results I found in each study and to what degree the findings were able to answer the questions postulated above. In addition, I examine the ramifications that the methodology has on author identification. I discuss how this study can benefit future author identification investigative endeavors as well as the potential limitations that might occur.

Finally, I propose ideas and suggestions for additional analyses in an effort to foster a consistent author identification methodology that when applied to a “forensic” problem will transcend myth and provide realistic expectations as to what linguists can offer under the rubric of forensic linguistics – author identification.

CHAPTER 2

LITERATURE REVIEW

Author Identification – Quantitative

Harold Love (2002, p. 12) cites Joseph Rudman's supposition that "every author has a verifiably unique style and to define authorship is to validate individuality" as the driving force behind the fascination linguists have with author attribution.

The search to verify unique individual styles with quantitative methodology has a long rich history in the field of Linguistics. These types of studies rely on a range of features that might serve as habitual characteristic markers. R.W. Bailey in "Authorship Attribution in a Forensic Setting" suggests that the features "...should be salient, structural, frequent, easily quantifiable and relatively immune from conscious control" (1979, p. 10). In order to be effective, the features must be measurable and countable, in some manner, to produce traits upon which we can distinguish differences among authors and validate authorship individuality.

History

D. I. Holmes in *Authorship Attribution* (1994) provides an exceptional historical review of the various criteria that have been considered for statistical authorship testing. Some of the most common criteria and techniques are outlined, and summarized below. The following criteria, techniques and authors are summarized as cited in Holmes (1994)

- **Word-Length** (as cited in Holmes, 1994, p. 88). Michael Oakes (p. 202) in his chapter “Literary Detective Work” states that the oldest surviving reference to the creation of a scientific stylometry was made in 1852 in a letter by de Morgan that the authenticity of some of the letters of St. Paul might be proved or disproved by a comparison of word length (1998). The St. Paul letters provide the bases for much of Christian theology. Mendenhall (1887) proposed that word length might be a distinguishing feature comparing Marlowe and Shakespeare. Brineger (1963) and Mosteller and Wallace (1964) tested the methodology in authorship studies and found that the context-dependence of vocabulary means that at best this methodology could only be an approximation. Later studies by Smith (1983) found that when genres are compared, they exceed any distinguishable features that may lead to author identity. Smith concludes that this methodology is so unreliable that any serious student of authorship should discard it from their repertoire.
- **Syllables** (as cited in Holmes, 1994, p. 88). Fuchs (1952) calculated the average number of syllables per word, the relative frequencies of the syllabled words, and their distribution in the text. He found that certain author traits correlated with the genre the author was using. A later study (1965) discovered syllable per word frequency distribution discriminated different languages more than individual authors. Bruno (1974) built on Fuchs work in a study examining heterogeneity within texts, reasoning that it serves as a sensitive stylistic marker. Brainerd (1974) followed suit by considering whether or not the number of syllables in a pair of consecutive words could be viewed of being independent of each other. His study found that independence could be obtained but that, as with the word-length studies,

changes in distribution were impacted by style changes – narrative versus conversation. He concludes, however, that further studies using this methodology may prove profitable.

- **Sentence Length** (as cited in Holmes, 1994, p. 89). Yule (1938) concluded that sentence length was not wholly reliable but raised important future questions concerning the definition of a sentence in quantifiable investigations. Further studies by Wale (1957), Morton (1965), and Sichel (1974), all resulted in similar conclusions. The consensus indicates that the disadvantage of using sentence length as a characteristic marker is that it is under the conscious control of the author and/or an editor. As with word length, Smith (1983) concludes that the information provided by this type of methodology is not sufficient as a stand-alone technique to discriminate between authors.
- **Distribution of Parts of Speech** (as cited in Holmes, 1994, p. 89). Sommers (1966) looked at the percentages of various parts of speech in text, suggesting intellectual habit could increase various usages and that empathy and attitude could be habitually expressed via verb usage. Antosch (1969) found that the verb-adjective ratio shows the ratio is highly dependent on the genre in which it is studied. For example, folk tales have a high ratio of verb-adjective use while scientific texts do not. Brainerd (1974) again found the ratio was not author specific, but genre specific. In a separate study (1973) of article/pronoun analysis of a novel versus a romance, he found that parts of speech were sensitive to variation in the degree of formality of the writing. Holmes suggests that the findings are hardly surprising and may provide interesting areas for future study.

- **Type-Token Ratio** (as cited in Holmes, 1994, p. 91). Tallentire (1973) focused on the measurement of richness and diversity of an author's vocabulary. A fundamental notion of author attribution is the assumption that a writer has an available amount of vocabulary, some of which the writer may prefer over others. Tallentire (1973) asserts that no potential parameter of style below or above that of the word is equally effective in establishing objective comparison between authors and their common linguistic heritage. A simple measure of N (the number of word (tokens) occurrences in a sample text divided by V (the number of vocabulary types) in the sample text equates to the type-token ratio R . Holmes points out that numerous studies indicate that this methodology is severely limited by the lack of stability of this equation in respect to sample size.
- **Yule's K Characteristic** (as cited in Holmes, 1994, p. 92). Yule (1944). The K value is another measure of vocabulary richness based on the assumption that the occurrence of a given word is based on chance. Bennett (1969) used it on common nouns in an investigation of vocabulary richness and found that within plays it has some value between specific acts, but none between actual plays. Tallentire found that the limitation of using it on just common nouns results in too wide a range of K values to be effective and any study should be expanded to include all words in a sample text. Sichel (1986) proved that under the distribution assumption, the occurrence of a word is based on chance and the property of K may have future investigative value.
- **Word Frequencies** (as cited in Holmes, 1994, p. 96). Zipf (1932) looked at how many times a word occurred in a given corpus and ranked the words in decreasing

frequency obtaining a straight line configuration resulting in Zipf's First Law.

Tallentire (1972) found difficulty with using word frequencies in authorship attribution because the relatively common set of highest frequency words recur in the same proportion in different authors as do the most common lowest frequency words.

- **Cusum (Cumulative Sum Charts) analysis.** Holmes (1998) in "The Evolution of Stylometry in Humanities Scholarship" also reviews the Cusum analysis. Despite overwhelming evidence that the technique is unproven, this type of analysis continues to surface in courtrooms around the world as an acceptable method of determining authorship. Proposed by Morton in 1990, the hypothesis of a cusum analysis is that each person has a unique set of habits they use when communicating, whether in spoken or written form, and that these habits are quantifiable. Morton relied most frequently on particular components of a person's sentences, for example (short words – defined as words of two or three letters, vowel words – defined as words beginning with a vowel, and a combination short + vowel words). Two plots are then generated and superimposed over each other - one plot for sentence lengths and one plot for the habit in question being investigated. Holmes explains, "The central premise for interpreting these cusum charts is that the two values (sentence length and the number of habit words per sentence) should parallel each other in the utterances of any one person" (p. 114).

A significant divergence between the two values would, Morton claims, demonstrate different authorship for part of the text since it would indicate a difference in rate of use of the habit. The technique was quickly adopted by defense attorneys to defuse and cast doubt about confessional statements made by their

clients. Skepticism quickly arose regarding the technique and independent studies by (Canter, 1992; de Haan and Schils, 1993; Hardcastle, 1993, 1997; Hilton and Homes, 1993; Sanford *et al.*, 1994; Holmes and Tweedie, 1995) investigated the method. Each study found the methodology unreliable. While proponents continue to put forth the technique, these studies question the subjectivity of the interpretation of the charts and that the underlying assumption regarding consistency is false. The studies indicate that a cusum should not be used as a definitive indicator of authorship.

All of these characteristics are still considered in the search for author identification; however, the success of various studies involving function words is now at the forefront of stylistics. It is these most commonly and frequently used words that may best personify language habit because variations in their use are often detectable, measurable and more importantly outside the realm of conscious thought.

Function Words

David L. Hoover (2001) explains, “because of their high frequencies in the English language and their low semantic load, the most frequent function words have long been assumed to lie outside the conscious control of authors” (p. 422). They are taken to reflect deeply ingrained linguistic habit. Hoover cites evidence from recent neurophysiological work by A. D. Friederici (1996) on the speed and location of the processing of closed-class (function words) versus open-class (context words) in the brain, which suggests that after about age ten speakers process close-class words (function) more rapidly and in a different area of the brain than open-class words (p. 422). He maintains that this ‘automatic’ processing of function words by

speakers tends to support the possibility of an author ‘word print’ that could be used to determine authorship.

Oakes in his chapter “Literary Detective Work” (1998, p. 199-248) credits A. A. Ellegård as the first to use function words in his study of the *Junius Letters*, political pamphlets written in 1769-1772 under the pseudonym *Junius*. Authorship of the letters had been attributed at various times to no fewer than forty people. The letters were comprised of approximately 150,000 words of text. Ellegård compiled an ordered list of words that he felt were positively or negatively characteristic of *Junius* in the sense that the author used them more or less frequently than his contemporaries of the same era. Ellegård then calculated a distinctiveness ratio, which divided the relative frequency of a specific word in *Junius* by the relative frequency of the word in a million-word sample of non-*Junius* writings. Ellegård’s study showed that Sir Philip Francis (the most probable contender for authorship of the *Junius Letters*) was the most likely candidate. Most notable about this study is that Ellegård did not use a computer for counting the words but relied solely on his intuition as to which words occurred more or less frequently.

In “Non-Traditional Authorship Attribution studies in Eighteenth Century Literature Stylistics Statistics and the Computer”, however, Joseph Rudman states that it is Mosteller and Wallace’s work on the twelve disputed *Federalist Papers*, utilizing function words, that are arguably the most famous and most successful study of author attribution. Rudman lists eighteen well-known subsequent studies that test Mosteller and Wallace results or use their methods and techniques in their own studies to show the influence this seminal work has had on author identification. He further maintains that almost every author attribution study that makes use of a computer, stylistics and statistics today, cite Mosteller and Wallace for one reason or another (Rudman, 2002).

The *Federalist Papers* (77 essays) were published anonymously in 1787-88 by Alexander Hamilton, John Jay, and James Madison to persuade the citizens of New York to ratify the constitution. Hamilton is the author of 43 papers, Jay of 5, and Madison of 14. Twelve of the essays remained in dispute between Hamilton and Madison. In his article “A Statistical Study of the Writing Styles of the Authors of *The Federalist Papers*”, Mosteller states, “Words offer many opportunities for discrimination. Some vary considerably in their *rates of use* from one paper to another by the same author; others show remarkable stability within an author. For discrimination purposes we need context-free or ‘function’ words to conduct reliable comparisons” (1987, p. 133). Wallace and Mosteller’s study showed that context words like ‘war’ suggest that the *rate of use* of the word depends on the topic under discussion. For example in discussions of the armed forces, the rate is expected to be high, in discussions of voting, low. He explains that he and Wallace “called words with such variable rates “contextual” and we regard them as dangerous for discrimination” (p. 134). By employing function words such as prepositions, conjunctions and articles as discriminators, and then applying Bayes’ theorem Wallace and Mosteller were able to determine probabilities of authorship to a high degree of accuracy.¹ Mosteller concludes this paper with his thoughts about function words, “We were surprised that, in the end, utterly mundane high-frequency function words did the best job. Though we love them for their lack of contextuality, their final strength was as unexpected as it was welcome” (p. 139). Remarking on authorship problems in general Mosteller extends the following insights in reference to function words usage in authorship studies:

¹ Bayes’ theorem is a statistical theorem describing how the conditional probability of a set of possible causes for a given observed event can be computed from the knowledge of probability of each cause and the conditional outcome of each cause (www.freedictionary.com).

1. “The Function words of the language appear to be a fertile source of discriminators, and luckily the high-frequency words are the strongest” (p. 140).
2. “Contextuality is a source of risk. For this reason it is important to have a variety of sources of material, to allow variability among the sources to emerge, and to give a basis for the elimination of words that show substantial heterogeneity” (p.140).
3. “Pronouns and auxiliary verbs appear to be dangerously contextual” (p. 140).

Wallace and Mosteller’s seminal work and the insights he discusses above, continue to be the foundation of function word studies today.

J. F. Burrow’s 1987 publication “Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style” appears to heed Mosteller’s remarks on authorship problems but states that his successful study, which uses the 30 most common word-types of Jane Austen’s six published novels, did not discern between function-words and content-words. The hierarchy was compiled and the words chosen based solely on highest frequency as he contends that the 30 most common word-types in many English texts make-up two-fifths of all the word-tokens used in them. Burrow’s further maintains, “If the patterning of the connectives testifies to the *ordonnance* of a style, that of the pronouns, verbs, and articles begins to flesh it out” (p. 62). Using a correlation matrix approach, Burrows was able to compare and contrast Jane Austen with other authors. His study distinguishes between not only difference in style between authors but also differences in style between Jane Austen at different stages of her career. Burrow’s subsequent work in the late 1980’s and early 1990’s continues to expand on the viability of function words as good discriminators in an author attribution study but with modifications.

Burrows' paper "An Ocean Where Each Kind...': Statistical Analysis and some Major Determinants of Literary Style" which deals with differences among Jane Austin's characters and differences between Jane Austin and other authors, he suggests that certain word types would need to be excluded. He explains that when there are marked differences between the comparative texts, it would be necessary to exclude certain word-types from the analysis, especially when indirect and direct speech is involved. Huge differences between pronouns and the inflected auxiliary verbs can tend to obscure all other resemblances and differences between texts and a comparison would have little meaning (1989).

By 1992, in "Not Unless You Ask Nicely: The Interpretative Nexus between Analysis and Information," Burrows modifies his function word count to the top most frequent function words. He states "the number of word-types tabulated is a matter of convenience: the top fifty, whatever they may be, make up about half of all the word-tokens in most texts written in English" (p. 92).

Burrows successful confirmation of Wallace and Mosteller's work regarding function words as viable author characteristic markers is now an established stylometric fact. However, more importantly, Burrows' analyses have served to regenerate an interest in stylometry as a viable tool for authorship investigation based on his quantitative computer methodology: multivariate analysis. Holmes (1998) suggests that this contribution is so phenomenal, he believes it should be ranked alongside the seminal function word work of Mosteller and Wallace (p. 113).

Multivariate Analysis

The advent of personal computing has allowed linguists to pick out less conspicuous characteristic features and count them far more quickly and accurately than was possible in earlier studies. With such advanced computer technology in the past two decades, author attribution linguists (Burrows, 1987, 1989, 1992; Holmes, 1994; Baayen and Tweedie, 1994) have had significant success with principal component analysis, which relies on the most commonly used (highest frequency) words in any text. Burrows, considered the father of this methodology, explains in *Computers and the Study of Literature*, “in this method of analysis, the word-types are allowed to “choose” themselves, to interrelate at their ‘choice’, and to show up whatever mutual patterning is most influential as an expression of resemblances and differences with a given set of texts” (1992a, p. 153). Holmes (1998, p. 114) emphasizes, “The trend towards usage of multivariate statistical methods is now so established in stylometry that it is unusual to find papers which do not use them.” Using principal component analysis, Holmes and Forsyth (1995) in their study “The *Federalist* revisited: New Directions in Authorship Attribution” confirm Mosteller and Wallace seminal 1964 findings that Madison is the probable author of the twelve disputed papers, corroborating that this multivariate technique can validate the Bayes’ theorem used in 1964. Tweedie, Holmes and Corns (1995) in “The Provenance of *De Doctrina Christiana*, attributed to John Milton: A Statistical Analysis” summarize what they refer to as the “Burrows methodology”, “... as a dimension-reduction technique that enables textual samples to be plotted in just two dimensions so that clustering and outliers are clearly visible” (1998, p. 78). They contend that the methodology is so powerful they can extend it into neo-Latin, an area and language where it has not been previously applied.

Baayan, Van Halteren and Tweedie in their study “Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution” used principal component analysis on experiments in author identification with a syntactically annotated corpus. Principal component measures that were previously applied to function words and their frequencies of use were applied in a similar manner to syntactic phrase rewrite rules in the annotated corpus. Using this methodology on the annotated corpus they found, “... that the use of function words for classification purposes is an economical way of tapping into the use of syntax, but that the direct examination of the frequencies of syntactic construction leads to a higher discriminatory resolution” (1996, p. 129). This study suggests that the frequencies with which rewrite rules are put to use might provide a better determiner of authorship than function word usage. In addition to their seminal findings using the methodology on syntax, they found that genre masks author specific variation on important principal components. Unfortunately, although the results were extremely promising, the lack of syntactically annotated texts continues to hinder future studies.

Increasingly, supportive multivariate analysis techniques such as cluster analysis and discriminant analysis have been applied to author attribution. In an analysis of the affinities (date, genre or authorship) of 100 plays by various authors from the Shakespearean period, Hugh Craig in “Is the Author Really Dead? An Empirical Study of Authorship in English Renaissance Drama” found that by using cluster analysis that authorship is a better explanation for clustering than genre or date. He states, “With multivariate statistical procedures one can track the crosscurrents...” (2000, p. 132). Holmes (1998, p. 114) explains that cluster analysis has been used in conjunction with principal component analysis in studies involving Mormon Scriptures by Holmes (1992), to examine Classical texts in Frischer *et al* (1996) and to consider eighteenth century texts by Mannion and Dixon (1997).

In a study of the plays of Thomas Middleton compared to his contemporaries in “Authorial Attribution and Computational Stylistics: If you can Tell Authors Apart, Have you Learned Anything about them”, Hugh Craig found that it is possible to achieve a good classification of Middleton’s work using discriminant analysis. Craig (1999) cites (Van de Geer, 1971, p. 243-72) explanation that “Discriminant analysis is a well-established multivariate technique which creates a function designed to separate pre-defined groups of observations, a function which can then be used to classify cases whose group membership is unknown” (p. 105). Holmes (1998, p. 114) explains that discriminant analysis has been also used to analyze Shakespeare (Ledger and Merriam, 1994) and the gospel of St. Luke (Mealand, 1995). These supportive results can only serve to strengthen conclusions originated from principal component analysis.

Despite significant strides using principal component analysis and other multivariate methods, Burrows in his 2003 article “Questions of Authorship: Attribution and Beyond” recently cautioned that “we have not yet reached, as Joseph Rudman (1998, 2000, p 170) would wish, fixed upon a single method of analysis or identified a verifiably unique style. Our methods are increasingly reliable, our use of them is ever more rigorous, and we have vast new corpora to strengthen our comparisons” (p. 7). Burrows explains that it is the form of the authorial problem that will dictate the main analytical tool to be applied. Discriminant Analysis is best used in open cases of author inquiry whereby the question is whether specimen X belongs to set A or B. He also contends that cluster analysis, when used in conjunction with principal component analysis, seems to yield the best corroborative results of potential authorship. Conversely, he also suggests, however, that as language analysts we do not forget “principal component analysis is *not intrinsically a test of authorship but only a comparative resemblance*”

(p. 8). Closed cases that search for the question whether Specimen X absolutely identifies with A or B require much more stringent testing before any absolute inference could be made, and in fact using quantitative statistical methodology it is highly unlikely absolute resolutions will ever be attained. He reminds us that “...these methods do not yield the “verifiably unique” stylistic signatures desired by Joseph Rudman or reach the “holy grail of stylometry” as David Holmes (1998, p. 116) describes it” (p. 26). He further prompts us that quantitative statistical multivariate analyses deal in probabilities, not certainties. As linguists interested in author attribution, multivariate analyses of potential authorship can lead us to determine with a high probability who might belong in Set A or Set B, but new techniques and advances will be required in conjunction with this methodology to deal in the realm of absolutes.

What is evident is that all quantitative authorship studies begin with a set of characteristics the researcher believes might yield the individually verifiable features that may provide author identification. No one factor works in all circumstances and there is no current consensus as to the best method. Any characteristic or a combination of characteristics may play a role in authorship identification, and the choice of approach is dictated by the particular study.

Author Identification – Forensic Linguistics

The majority of linguists who are involved in large quantitative studies of authorship identification do not refer to their studies as forensic. Forensic Linguistics is a branch within Linguistics whereby the language studied provides scientific evidentiary analyses involving language, crime and the law, and these analyses are a more recent phenomenon. Jan Svartvik, who analyzed the statements of Timothy John Evans, hanged for the murder of his wife and baby, and posthumously pardoned, coined the term *forensic linguistics* in 1968 (Svartvik). The

term, however, lay dormant for years. The 1980's saw a resurgence of the term as systematic studies by linguistics emerged that could provide analyses specific to the legal arena.

McMenamin (2002, p. 67) in *Forensic Linguistics: Advances in Forensic Style* cites the early works of Danet (1980) on language of fact-oriented disputes, O'Barr (1982) on courtroom language, and Shuy's (1984, 1986) linguistic courtroom applications in discourse analysis, and claims these as a factor in the regeneration of the rapidly growing sub-field of applied linguistics.

Classification of the sub-fields that fall under the rubric of forensic linguistics is ever evolving including the scientific application of linguistics to trademark violations, language crimes such as threats and bribery, and the best known, the elusive search for the linguistic fingerprint or linguistic DNA that will irrevocably identify the perpetrator in a "whodunit" scenario.

Statements such as the following certainly incites the imagination and feeds today's forensic-crazed environment "The scientific analysis of a text – how mind and a hand conspire to commit acts of writing – can reveal features as sharp and telling as anything this side of fingerprints and DNA. Although we disguise our writing voice, it can never be fully masked. After the crime, the words remain. Like fingerprints and DNA" (Foster, 2000, p. 4). The concept of the linguistic fingerprint, that the difference between people's language use can be observed as easily as a fingerprint, however, is one that despite strides in the field of forensic linguistics remains unproven. Although the notion has been highly touted in the media, journals, books, and even written about as fact, minimal replicable data exists to substantiate the claim.

John Olsson in *Forensic Linguistics: An Introduction to Language, Crime and the Law* calls the concept "The myth of the linguistic fingerprint" citing numerous references to a "linguistic fingerprint" of which none has ever come to fruition. He suggests perhaps it is the

word ‘forensic’ that is responsible for the unproven concept as the fact that it collocates with words like ‘expert’ and ‘science’ cannot help but raise expectations (2004, p. 31).

Most forensic authorship problems involve a limited amount of text that does not lend itself to a purely quantitative analysis. The methodology in forensic linguistics expands to utilize qualitative methods that focus on features of the text that can be described as being characteristic of the author, described by McMenamin (2000) as stylistics. McMenamin cautions that in a DNA – linguistics analogy, DNA is a chemical and biological system, described in a natural paradigm; however, written language is a neurological, psychological and sociocultural system, described principally within a social-science framework” and that a DNA analogy creates expectations that can only sometimes be met (p. 59).

The fingerprint/DNA controversy aside, author identification utilizing stylistics can provide evidentiary analyses that suggest potential authorship. Stylistics investigation involves feature characteristics such as text format, punctuation, numbers and symbols, abbreviations, spelling, word formation, syntax, errors and corrections, high frequency words, and basic statistics. McMenamin (2000) provides a substantial list of style markers he found useful in eighty authorship identification cases. Stylistics also involves genre and style. Olsson (2004) explains that an individual author writes in an identifiable style but can also write in a genre that indicates a badge of membership. This can lead to a conflict as some features may be under individual control and some aspects of style related to the genre. A stylistic study must consider this issue when applying style markers to determine authorship.

Olsson used these types of markers in his “Dog Club Treasurer” case where the president and other senior committee members of a mid-west dog club had received a series of vicious anonymous letters. Since the details of the texts included club management and policy, and an

intimacy with virtually all of the committee members, he concluded the writer was probably a member of the committee. By examining, orthography, grammar confusion, spurious capitalization, and salutations, and comparing them to known committee member texts, he successfully identified the writer of the vicious letters (2004, p. 51-58).

Don Foster (2000) successfully identified Joe Klein, as the author of *Primary Colors* (used as a case study in this dissertation) by comparing the adjectival and adverbial vocabulary characteristics of Anonymous and Klein (p. 64-69).

From 1982 to 2000, McMenamin was involved in 204 cases that required a forensic linguistic analysis, of which 93% were author identification cases requiring a stylistic approach (2000, p. 207–231). Unfortunately, however, less than 10% of the cases were adjudicated and therein lies the quandary forensic linguists using stylistics for author identification currently face in the United States in the legal arena.

In the United States to determine whether a theory or technique will qualify as the basis of admissible evidence under rule 702 of the Federal Rules of Evidence, *Daubert v. Merrell Dow Pharmaceuticals* (1993) proposes key questions be answered but “the inquiry is flexible”.

- Can the theory or technique be tested?
- Has it been tested?
- Has the theory or technique been subjected to peer review and publication?
- Is the known or potential rate of error of the particular scientific technique known?
- Can a relevant scientific community be identified?
- Is there an expressed degree of acceptance of the theory or technique within the community?

As this relates to linguistic stylistics, the answers are yes, but with substantial qualifications. Theory and testing is continuous and Carol Chaski (2001) attempted the first significant study of error rates, albeit with mixed success, in a forensic context in her study “Empirical evaluations of language-based identification techniques”. In an attempt to show why testing of documents does not hold up in the U.S. courts of law, she applied several basic quantitative tests on controlled text sample writings. She found the majority of her quantitative tests did not establish any acceptable degree of probability. Linguists McMenamin (2001) and Grant and Baker, (2001) criticized her sharply for the style markers she chose but neither considered that sample size might have affected her results. Each corpus that Chaski developed had fewer than 1000 words. Baayen (2001) in *Word Frequency Distribution* cautions us that sample size is critical in any statistical analysis of language data, as the results can vary dramatically and in most forensic cases that involve language the first hurdle is generally sample size. For example, in the tragic JonBenét Ramsey case, the ransom note left at the scene contained only 317 words.

Currently, *U.S. v. Van Wyk* is cited as a case that demonstrates the weakness of forensic stylistics. The court objected to the credentials of the witness and rejected the testimony as to his conclusion of author identity; however, it is important to note that the court did admit the witness testimony regarding the comparison characteristics and markers between writings known to have been authored by the defendant and the writings in which authorship was questioned or unknown. The latter testimony has been allowed in one other recent case, *U.S. v. Spring* (2001) (McMenamin, 2002).

Replicable, verifiable data is critical to the success of forensic linguistic stylistics methodology. Issues regarding linguistic stylistic approaches in a legal environment will have to

be resolved before a court of law will consider these types of analyses scientifically sound and allow them into evidence and the following case studies illustrate the courts concerns.

In their comprehensive article, “The Murder of JonBenét Ramsey” published in *Crime Magazine*, J. J. Maloney and J. Patrick O’Connor recount the brutal murder of 6 year-old JonBenét Ramsey on Christmas night in 1996 that shocked America. They write that Don Foster, who had unmasked the anonymous author of *Primary Colors*, was contacted by investigators in early 1997 and given access to text written by suspects for comparison with the ransom letter left at the crime scene. Don Foster flew to Boulder in March of 1998 to brief investigators and representatives of the D.A.’s office of his findings regarding who wrote the ransom letter. Foster stated, “We can’t falsify who we are... Sentence structure, word usage and identifying features can be a signature.” Foster further opined “... it is not possible that any individual except Patsy Ramsey wrote the ransom note” (1999, no p. # listed).

Conversely, Gerald McMenamin, contacted by attorneys for John and Patricia Ramsey in 1997 and given direct access to both parties, conducted a similar study using stylistic methodology. McMenamin states in chapter 10 – “Case Study: JonBenét Ramsey” in *Forensic Linguistics: Advances in Forensic Stylistics*, “Patricia Ramsey is excluded as the writer of the questioned ransom letter.” He bases his conclusion on three facts: substantial dissimilarities between her range of variation and the ransom letter, that limitations of the available data did not diminish the significance of reliable data or indicate language disguise, and the range of variation measured in the question letter constitutes sufficient basis for comparison (2002, p. 181-204). This tragic case remains unresolved.

There is no question that in hindsight (i.e. The Unabomber case) qualitative stylistic studies using basic statistical methodology and stylistics can confirm authorship to high

probability. However, forensic linguistic analyses on current unresolved cases, such as the JonBenét Ramsey case, whereby forensic linguists come to completely different conclusions using the stylistic approach only serve to highlight to the courts the need for continuous testing before they will consider stylistics a viable option regarding author identification.

Terminology: Register - Genre – Text Type

The terms ‘register’, ‘genre’, and ‘text type’, have been used in many different ways by various disciplines and researchers.

In an “An Analytical Framework for Register Studies” Douglas Biber states “Most researchers agree in using ‘register’ to refer to situationally defined varieties as opposed to ‘dialect’, which refers to varieties associated with different speaker groups” (1995, p. 51). In other words, people speak or write differently depending on the occasion of use, and on their intended audience. I speak differently to my professor at the University than I speak to my friends in the coffee shop. I write differently in an email to a family member than in a research paper. He explains that beyond this rather general use of ‘register’ as a situational variety there is little consensus among researchers. Some use it as a catchall situational cover term with no discussion of the level of generality. Others discard it all together stating it has been used indiscriminately to cover a myriad of varieties of language and rhetoricians have generally used ‘genre’ instead of ‘register’ (1988).

Literary definitions of ‘genre’ feature literary conventions. In the “All American Glossary of Literary Terms”, Mark Canada defines ‘genre’ as a type of literature, a poem, novel, story, that belongs to a group of works that shares at least a few conventions or standard characteristics. For example, the Gothic genre often features supernatural elements, attempts to

horrify the reader, and includes dark, foreboding settings. Edgar Allan Poe's short story "The Fall of the House of Usher" belongs to the Gothic genre. He explains, "An understanding of genre is useful because it helps us to see how an author adopts, subverts, or transcends the standard practices that other authors have developed" (2006). Dividing literary works into genres is also a way of classifying them into particular categories. For example, a particular work is classified into fiction or non-fiction, and then classified into a category that will specify the form of the fiction, for example, drama. Fiction can be classified according to technique, fiction and short stories, or classified according to content and theme, fiction and mystery.

Biber in his essay "An analytical framework for register studies" in *Sociolinguistic Perspectives on Register* states that the word 'genre' was also used in the IPrA Survey of Research in Progress (Nuyts, 1988) as a language variety, along with jargon, argot, and slang. Examples of registers in this type of framework are journalese, legalese, and aviation language. This survey identified two major discourse types, conversation types and text types. Text types include advertisement, essay, joke, letter, literature (1994, p. 52).

In addition to the above, 'genre' and 'text type' were used in multidimensional studies by Biber (1988, 1989, and Biber and Edward Finegan (1989). Biber (1988) in *Variations across Speech and Writing* defines 'genre', "to refer to text categorizations assigned on the basis of external criteria relating to author/speaker purpose" (p.69). For example, the genre of an academic article on Asian history represents formal, academic exposition in terms of the author's purpose. In contrast, however, Biber defines 'text type', "as a grouping of texts that are similar with respect to their linguistic form, irrespective of genre categories" (p. 69). In the above example on the article on Asian history, while it represents formal, academic exposition (genre), "its linguistic form might be narrative-like and more similar to some types of fiction than to

scientific or engineering academic articles” (p. 70). The Asian history article would have a genre of academic exposition, but a text type that may be academic narrative. In essence, texts are similar with respect to their linguistic characteristics (syntactic, morphological, lexical) and types are similar with respect to their linguistic characteristics (Biber, 1989). This definition of ‘genre’ and ‘text type’ will prevail in the quantitative cross-genre author identification study in this dissertation. Many studies continue to interchange the terms ‘register’ and ‘genre’ and that will be highlighted when they occur.

Genre

Charles Ferguson in his essay “Dialect, Register and Genre: Working Assumptions About Conventionalization” in *Sociolinguistic Perspectives on Register* explains it was noticed early in human history that people differ in their speech and writing depending on where they come from and where they belong in society, we refer to this as dialect. Another variation is that language differs with different occasions of use that comprises our different registers. The analysis of different kinds of literary texts, including their structure and uses is referred to as genre studies and goes back to Aristotle’s *Poetics*. He maintains, “... after WWII many scholars tended to deemphasize, neglect or deny the usefulness or even the possibility of genre studies” (1994, p. 17). He further explains that the 1970’s saw a resurgence of genre analysis and genre theory as it became the focus of many literary research studies. A general recognition developed along with the resurgence that genres in the sense of discourse types exist just as much in nonliterary spoken or written texts as they do in literary texts. Some of the more insightful studies on genre and the principles behind their emergence came from linguists and discourse analysts. Ferguson explains that in 1983, Brown and Yule produced a sociolinguistically

oriented standard textbook on discourse analysis that gives an indication of the accepted but unanalyzed status of genre in the field and twenty-five examples are offered in the book in connection with various observations about discourse analysis. Examples are chat, love-letter, news broadcast. Ferguson also discusses Linguist Paul Freidrich's 1988 work that studied the literary genre - the sonnet. His study of the sonnets history, variant forms and present status shows the formal variants in a genre that is regarded as highly specified and also raises the issue of genres moving from one speech community to another (1994, p. 20-23) .

Biber and Finegan in their article "Drift and the Evolution of English Style" follow three English-language genres of written prose from the seventeenth century to the twentieth: fiction, essays and personal letters of literary figures. Their multidimensional approach describes the discovery of a general, overall long-term drift in English style over four centuries (1989).

Genre – Author Identification

Genre as we saw in the various quantitative methodologies that are applied to author identification appears to play a critical role in the results.

- Smith (1983) - word length studies - found that when genres are compared, they exceed any distinguishable features that may lead to author identity.
- W. Fuchs (1952) - syllable studies - found that certain author traits correlated with the genre the author was using. Brainerd (1974) followed suit by considering whether or not the number of syllables in a pair of consecutive words could be viewed of being independent of each other. His study found that independence could be obtained but that as with the word-length studies changes in distribution were impacted by style changes – narrative versus conversation.

- Sommers (1966), Brainerd (1974) – distribution of parts of speech studies – found the ratio verb-adjective ratio was not author specific but genre specific.
- Olsson (2004) – stylistic studies - explains that an individual author writes in an identifiable style but can also write in a genre that indicates a badge of membership. This can lead to a conflict as some features may be under individual control and some aspects of style relate to the genre.

In their study using principal component analysis on syntactically annotated corpora, Baayan, Van Halteren and Tweedie (1996) found that genre masks author specific variation on the most important principal components. They state, “Apart from the advantages of syntax-based methods, our analyses have also shown the need for closer examination of the relative importance of register-specific and author-specific variation” (1996, p. 129). Of special concern to Baayan *et al.* were Innes and Stewart, who is one and the same person. Stewart (1963) writes in the genre of literary criticism but uses the pseudonym Innes (1966) for writing in the genre of crime fiction. Their initial pilot study clearly indicates that, “for one author differences in register can be stronger than differences within a register between texts of different authors” (p. 122). Stewart’s literary criticism texts clustered much more closely with scientific texts than with his own novel. They decided to opt for a controlled experiment on authorship attribution within a single register (genre).

José Binongo confirms Baayan *et al.*’s findings regarding the influence of genre in a pseudonymic environment with his study of Joaquin, a writer notorious for his hyper-complex sentence constructions decades ago, who now writes children’s books in a much less complex manner. In his article in “Joaquin’s Joaquinnesquerie, Joaquinnesquerie’s Joaquin: A Statistical Expression of a Filipino Writer’s Style” Binongo explains that common stylistic parameters, that

depend on genre or intended audience, can blur stylometric investigations on authors with varying styles. Joaquin's earlier writings initially clustered more closely with four works of his contemporaries than with his own recent children's writings. However, Binongo's initial findings also found that cross-genre concerns might be overcome using a principal component analysis of high frequency words whereby a change in writing style does not impact the stability of the highest frequency functions words. He suggests that further experimentation needs to confirm this result (1994).

Genre – Multidimensional Studies

Baayen *et al.* (1996) suggest, in light of the success of their experiment, that multi-register (genre) corpora such as those used by Biber in studies of register variation (genre) are promising for questions of inter-register (genre) authorship attribution. Biber's seminal work, *Variation across Speech and Writing* attempts to identify a range of features, which characterize written and spoken speech. He explains that texts can be related along particular situational or functional parameters such as formal/informal, literary/colloquial. He states, "These parameters can be considered dimensions because they define continuums of variation rather than discrete poles" (1988, p. 9). Biber operates on the premise that there are few absolute differences between speech and writing and that there is no single parameter that will distinguish speech and writing. For example, although it may sound awkward, a formal sentence that is written can be expressed orally, and conversely a sentence that is spoken can be used as a written form and retain its meaning and understandability. He seeks to "systematically describe the linguistic characteristics of the range of genres in English, whether typically spoken, typically written, or other" (p. 55). The study attempts to identify how genres vary based on linguistics parameters.

He designed the study so that “any individual genre can be located within ‘oral’ and ‘literate’ space, specifying both the nature and the extent of the differences and similarities between that genre and the range of other genres in English” (p.55).

Biber explains that the framework for this process includes textual dimensions and textual relations whereby “Dimensions are bundles of linguistic features that co-occur in texts because they work together to mark some underlying function. Relations are defined in terms of dimensions; they specify the ways in which any two genres are linguistically similar and the extent to which they are similar” (p. 55). Using a multivariate analysis on 67 linguistic features, Biber identified seven dimensions. He then examined fifteen written genres from the LOB Corpus and six spoken genres from the London-Lund corpus, and two types of American letters, which associated with each of these dimensions and determined how they were related linguistically and to what extent.

Biber developed a clear example of how this process affects and determines genre. He takes four texts, scientific text, panel discussion, conversation, and fiction. After examining the texts from a linguistic perspective, he identifies two co-occurrence patterns that belong to two independent dimensions – many passives and nominalizations and few passives and many nominalizations (1988, p. 9-27). The dimensions can be then be plotted to illustrate their independent status among the texts as demonstrated below:

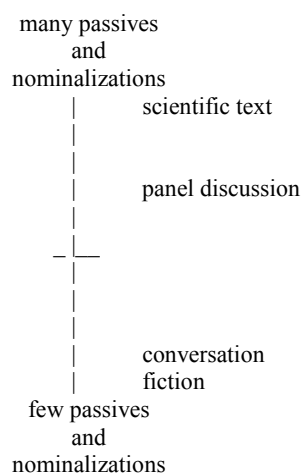


Figure 2.1a One-dimensional plot of four genres (Biber, 1988)

Other linguistics dimensions can further co-occur independent of the first pattern, resulting in a two-dimensional plot, hence the term ‘multidimensional’ studies. For example:

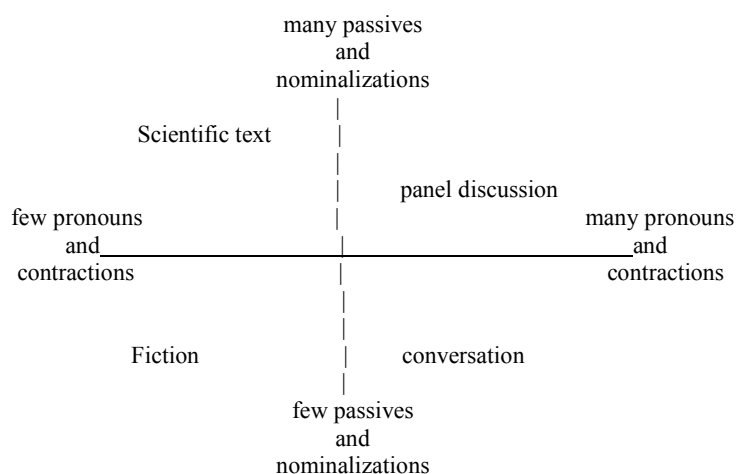


Figure 2.1b Two-dimensional plot of four genres (Biber, 1988)

Once plotted, Biber investigates why these particular sets of features co-occur in text and what parameters relate that will influence their systematic use across a range of texts and genres.

Biber's seminal multidimensional studies using cross-genre corpora provides those investigating authorship with a solid foundation of what linguistic features predicated on genre a study must consider, what features might cloud or mask identity, and to what extent genre can influence the author. For example, as the above plot shows the use of few passives and nominalization features in fiction and conversation, and conversely the use of many of those features in scientific text and panel discussions. The following two studies on *Psychic language* and *Genre* utilize many of the concepts and methodologies discussed in this chapter. Both studies involve author attribution and use multivariate analyses on high frequency function words. While the study on *Psychic language* seeks to determine whose voice is captured in a particular text, the *Genre* study strives to determine if the influence of genre between fiction and non-fiction is strong enough to mask the identity of a pseudonymic author. The extraordinary studies described in this chapter allow me to postulate that probable author identification is linguistically possible and provide me with the solid foundation to pursue my own investigations.

CHAPTER 3

METHODOLOGY

An overview of the methodology used in both of the following studies is outlined in this chapter. I begin both studies by building specific corpora for all the texts involved in the studies using readily available electronic data or converting printed texts into electronic data using an OCR (Optical Character Recognition) program. After establishing each corpus, I will use WordSmith Tools to determine word frequency usage, Excel to format the data, and SPSS for statistical data analysis. My initial analysis of the data will start with principal component analysis (PCA) referred to as the ‘Burrows method’, which D.I. Holmes called, “the first port-of-call for attributional problems in stylometry” (1998. p. 115). PCA will help determine how the function words behave and which ones reflect characteristic markers for individual authors. Once applied, discriminant analysis (DA) will be employed to classify the authors into groups based on the measurements of the individual word frequency usage. Lastly, cluster analysis (CA) will be utilized to reclassify authors into uniquely defined subgroups in a secondary procedure to confirm my DA findings. These analyses will determine whose voice is found in *The World Beyond* in the psychic study as well as determine if Joe Klein retains his characteristic habitual markers when crossing from the genre of fiction in *Primary Colors* to the genre of his non-fiction from his *Newsweek* column political writings “Public Lives”.

Corpora

Critical to my analyses is that I develop relevant corpora. A corpus is a compilation of running text derived from naturally occurring language, either written or spoken, upon which we can do linguistic analysis. Corpora allow us to focus on what is probable in language using statistical tests with which we can make inferences about the features of the language we are analyzing. John Sinclair (1991) in *Corpus, Concordance, Collocation* explains that a well-built corpus shows us that the words we choose are not random and that they are highly patterned (p. 110). The usages of these words become habitual to the individual and reflect characteristic markers from which patterns emerge. A good corpus allows us easily to retrieve and analyze these habitual characteristic markers. For example, we might wish to determine what words in a text collocate and how are they used by a specific author. Prior to the advent of the technology that allows construction of a relevant corpus of naturally occurring language, it would not have been possible to do the large-scale quantitative comparisons.

Optical Character Recognition (OCR)

If a text is not already in an electronic form, it has to be keyboarded into a software word processing program like Notepad or Microsoft Word or converted electronically by a special OCR software package using a scanner. OCR is a process that converts paper documents, or in the case of parts of these two studies entire books, into editable, manageable electronic data upon which further analysis can be done.

AIM, the Association for Automatic Identification and Data Capture Technologies explains that engineering attempts at automated recognition of printed characters started prior to

WWII. Unfortunately the process lacked value until the mid 50's when Banks and Financial Institutions identified that development of this type of technology would greatly benefit check processing, which had become the single largest paper processing application in the world. The check processing industry eventually chose a technologically more advanced but unrelated process called Magnetic Ink Recognition (MICR) instead of OCR. OCR was still in its infancy at the time and simply did not perform as well as the more advanced technology provided by MICR for check processing (AIM, 2000. p. 3).

Subsequent generations of OCR equipment and software saw new technologies applied to the process identifying that current packages are now powerful tools for reading and converting printed data such as books into manageable, analyzable, electronic data. Susan Hockey in *Creating and Acquiring Electronic Texts* provides a clear explanation of how OCR systems work. She explains, "OCR systems work by making an image representation of a page of text on a scanner and then attempt to convert the image or picture of the page into text characters by recognizing each character" (2000, p. 21). Once converted into text characters it can be electronically edited and analyzed. The only issue at hand is that any material prepared via OCR for analysis must be carefully proofread for accuracy. Unfortunately, OCR to date has not achieved a 100% read perfection rate. Even a 99.9% accuracy rate would mean one error per thousand characters or about every 10 – 15 lines. However, the most recent OCR systems are very forgiving of the printing quality found in printed materials and provide a suitable process for the conversion of the books and articles needed in these two studies into electronic data.

Most scanners come with a limited version of an OCR software program. I initially began my OCR process with a limited version; however, the magnitude of the amount data that I was scanning required that I upgrade to a professional version. My professional version allows

me to spell check the text for anomalies as I am scanning each page. In addition, the upgraded version significantly decreases the time it takes to scan each page. Initially the scanning and conversion of each page took approximately one minute and did not recognize that a book laid flat on a scanner was two separate pages. The upgraded version allows for a split screen scan, which determines a book is being scanned and that each page is an individual entity. The scanning time decreased to less than 23 seconds per open book page. An average book takes approximately 10 hours of continuous scanning to complete. This is an incredible time saving process considering the alternative would be to keyboard the entire text manually. Once scanned into electronic format I am able to save the data to Microsoft word for further editing.

Recent OCR systems are very forgiving of printed materials. The upgraded version greatly increased the accuracy of the data scanned; however, some consistent read perfection errors occurred. I found the professional version I used on all of my texts had difficulty appropriately recognizing words within texts that were italicized or bolded. Em-dashes also presented a concern, and although the OCR software does allow editing once scanned, saving the text to Microsoft word allows for much better editing options, whereby em-dashes, spelling errors and recognition concerns can be handled with greater ease. As I was scanning typed print, I found my texts were being converted with 99%+ accuracy. For example, the entire text of *A World Beyond* (58,265 words) had 147 recognition errors that required review and change (.0025%).

Once the text was converted and saved into Microsoft Word, it was checked for accuracy. I eliminated page numbers, reversed bolded and italicized words into regular text, and removed non-word elements such as em-dashes from the texts. Each text was then saved as a series of

text files in approximately 5000 word segments removing all formatting, which facilitates its being read by the software program WordSmith Tools.

WordSmith Tools

Mike Scott, the programmer responsible for WordSmith Tools, explains, “Oxford WordSmith Tools is an integrated suite of programs for looking at how words behave in a corpus. The WordList tool lets you see a list of all the words or word-clusters in a text, set out in alphabetical or frequency order. The concordancer, Concord, gives you a chance to see any word or phrase in context -- so that you can see what sort of company it keeps. With KeyWords you can find the key words in a text. The tools are used by Oxford University Press for their own lexicographic work in preparing dictionaries, by language teachers and students, and by researchers investigating language patterns in lots of different languages in many countries world-wide”(2004). WordSmith Tools was used specifically in these two studies to determine the word frequency usage of each developed corpus in 5000 word segments, and to establish the word hierarchy that was to be used as variables for statistical analysis using SPSS statistical software when all of the corpora were compiled. First, a compilation of all the individual corpora to be used in the study are combined into a master corpus and read by WordSmith tools to ascertain the word hierarchy (50 most frequent function words) that is to be studied. Then each individual corpus is read by WordSmith Tools to ascertain the exact word segment length, then to search each word in the hierarchy list to determine the number of times it is was used. The ensuring results are plotted into an Excel spreadsheet as raw data.

Excel

I used the Excel spreadsheet program to develop the matrix of functions words used for each author. Excel imports easily into SPSS for analysis. In developing a matrix that can be imported, it is imperative that the data is formatted in a manner that can be read by SPSS. For this type of study, the variables (function words) are represented in rows. Authors, individual segmentation, and their usage of each word are represented in columns as detailed below.

Table 3.1 Excel author segmentation – raw data sample

WORD RANK			1	2	3	4	
Author	Texts	Segment	THE	OF	AND	TO	Segment Length
		1	241	122	137	149	5041
		2	239	148	127	151	5030
		3	134	88	133	78	2992

Once the raw data is input, it is easier to convert the raw data, number of times a word is used by a specific author, into relative frequency data in Excel than in SPSS. Raw data must be converted into relative frequencies to obtain accurate statistical results. The raw data must be copied into another Excel sheet and the cells formulated to be divided by the true Segment Length as shown below.

Table 3.2 Excel author segmentation – relative frequency sample

Author	Texts	Segment	THE	OF	AND	TO	THAT	Segment Length
		1	0.047808	0.0242015	0.0271771	0.0295576	0.016465	5041
		2	0.0475149	0.0294235	0.0252485	0.0300199	0.0151093	5030
		3	0.0447861	0.0294118	0.0444519	0.0260695	0.0177139	2992

SPSS 13.0

SPSS 13.0 is a comprehensive statistical software package for analyzing data. SPSS (Special package for Social Sciences) is among the most widely used software programs for statistical analysis in social science. It was used specifically in these studies to perform multivariate statistical analysis to determine authorship based on function word usage.

Unfortunately, SPSS is not as user friendly as some other software programs; however, if you format the SPSS variable view to match the Excel spreadsheet identically, it will easily import your data for analysis.

Multivariate Statistical Analysis

Multivariate approaches to author attribution were introduced in the late 1980's and the early 90's by John Burrows as discussed in Chapter 2. His significant findings resulted in bringing multivariate methodology to the forefront of author attribution and it is now considered the standard for solving authorship problems quantitatively. The multivariate statistical procedures chosen for the studies in this dissertation were predicated on the methodology used in earlier literary studies of authorship attribution and their ensuing successful results. As I am a linguist not a statistician, I initially consulted with the University of Georgia Statistical Consulting Center to determine which multivariate methodology to apply to my data. Dr. Dan Hall and Graduate Student Jing (Jasper) Xu met with me to determine the best matrix arrangement for my data in the psychic language study. It is important that when applying multivariate methodology that you have at minimum as many cases as you have variables. Cases are considered the author's individual text segments. Upon completion of my data matrix, they guided me through the principal component and discriminant analysis process using an alternate

statistical package (SAS) to which I did not have access. Further research resulted in my being able to replicate the initial principal component analysis results using SPSS, to expand my study to employ discriminant analysis for author identification and include cluster analysis as a secondary confirmation of my findings. Three books were invaluable in my research regarding multivariate analysis and the procedures involved in applying the specific analyses: Marija J. Norušis, *SPSS 13.0 Statistical Companion* (2005); Dallas E. Johnson, *Applied Multivariate Methods for Data Analysts* (1998); Samuel B. Green, Neil J. Salkind, *Using SPSS for Window and Macintosh* (2003). The same three multivariate techniques (principal component analysis, discriminant analysis, and cluster analysis) initiated in the Psychic language study are applied to the Joe Klein Genre study. I refer to the compilation of these three techniques as the multivariate triad.

Principal Component Analysis (PCA)

Norušis explains that Factor analysis allows one to reduce a large number of correlated variables (in these studies the 50 most frequent function words) to a more manageable number of components (factors) that can be used for further analysis. Principle component analysis (PCA) is considered the simplest method. The first principal component accounts for the largest amount of variance among the data, the second principal component accounts for the next largest amount of variance among the data and is uncorrelated to the first. Once you have extracted a small number of components (factors) you can observe that some variables correlate more highly with some components (factors) than with others. You can then use the correlation patterns between the components (factors) and the variables to interpret the components (2005). For example, consider that we have five authors and fifty high frequency function words. Which

words became characteristic markers that indicate potential habitual underlying use? Do the parts-of-speech that the function words represent (e.g. prepositions, verbs,) play a role in determining what accounts for the greatest variance among the components (factors). The more manageable number of components is also used in additional statistical techniques such as discriminant analysis rather than the initial variables (50 most frequent function words).

In addition, as I was learning to apply PCA in SPSS on my data, I encountered a question that need to be considered by others who may wish to replicate my study. A rescaling of the variables generates principle components. Many researchers using PCA apply the correlation matrix to their data. Mainly because it tends to be the default on most statistical packages such as SPSS or in some statistical packages the only option. The consultants I initially worked with applied the covariance matrix. Further research indicates that either option can be used. Norušis, Green and Salkind do not address the issue and simply assume correlation, while Johnson advises that using correlation is equivalent to applying PCA to Z scores or standardizing the data rather than applying raw data values. She maintains that this assumption should not be made arbitrarily but that it is an unfortunate by-product of most statistic programs (1998 p. 108). For the purposes of this study, I chose to follow Norušis, and Green and Salkind's guidance. I did run the Psychic study using both correlation and covariance. The resulting generated components (factors) from PCA changed the discriminant analysis results and indicated that using correlation reflected a more conservative result than using the covariance matrix.

Discriminant Analysis (DA)

Discriminant analysis (DA) is a technique that is used to examine whether two or more mutually exclusive groups can be distinguished from each other. It can classify individuals into

one or more uniquely defined populations. To use discriminant analysis, you must have information for a set of cases of membership that you know. For example, consider that we have known texts (cases) from several authors. For each case, we have variables that are useful for distinguishing the groups – specifically we have the variables that we ran through PCA (fifty most frequent function words) from which we derived a more manageable number of components upon which we can do further analysis. Then consider we have a new text (cases) that we know came from one of our established authors but we are not sure exactly which one. Using the components derived from PCA, DA calculates discriminant scores that assign the unknown texts (cases) to a predicted author group. The objective of DA is to classify or predict which author most likely wrote the unidentified text (Green, 2003; Johnson, 1998; Norušis, 2005).

Cluster Analysis (CA)

Cluster analysis (CA) is similar to discriminant analysis in that it is a statistical technique that attempts to classify individuals into one or more subgroups. The major difference is in that cluster analysis you want to form groups based on the characteristics of the cases. In essence, an analyst does not know who or what belongs to what group. The overall goal is to identify the actual groups. For example, consider that we have segments of texts from various authors and we are unsure which segments belong to what authors based on the usage of the components developed in PCA, which determine how the cases might be similar. We start out with a number of cases that we want to subdivide into homogeneous groups. The algorithm of cluster analysis begins with each case being a cluster unto itself and at each successive step, similar clusters are merged until no further hierarchal clusters can be derived. Norušis maintains that unlike PCA

and DA, the benefit of CA is that we often don't have to make any assumptions about the underlying distribution of the data to apply CA (Green, 2003; Johnson, 1998; Norušis, 2005).

The methodologies described in this chapter serve to provide the foundation upon which each of the following studies is based. My goal is to insure that linguists can replicate with ease the processes and specific methodologies used in this dissertation.

CHAPTER 4

PSYCHIC LANGUAGE

In 1971, Ruth Montgomery, a renowned psychic, claimed that her deceased friend, sensitive Arthur Ford, was continuing to communicate with her in the form of automatic writing. Montgomery's glossary defines automatic writing as the production of written messages on paper or other surfaces, seemingly without the conscious thought of a living person. Shortly after his death and the abrupt cancellation of an alternate Montgomery manuscript, Montgomery claims that Arthur Ford communicated to her during her automatic writing process that he wished her to write a book in collaboration with him about life in the next stage of eternal life. The supposed collaboration resulted in the book *The World Beyond*; and Montgomery states, "This book I believe to be Arthur Ford's own account of the life in the next stages of existence beyond the portal man calls death" (1971, p. 3).

Problem Statement

Any judgment regarding the validity of this claim is purely faith based as generally the areas found under the rubric of the parapsychology industry are based on personal beliefs and values that cannot be scientifically corroborated. While believers maintain the results do not need science to legitimize the claims, skeptics are concerned that the lack of empirical authenticity for those claims allows those who make them to bilk millions of people out of their well-earned money and that the claims are little more than well played scams. Numerous texts

exist from psychics that like the one by Ruth Montgomery claim some or all of the written language in their texts is not their own but belongs to an entity speaking through them. No scientific studies regarding authorship have been attempted. I contend that a quantitative analysis of the language involved can determine if alternate authors exist within these texts.

Background of the Study

Unlike many psychic claims, which rely solely on one's own personal belief and value systems, I argue that this particular type of claim provides us with concrete evidence upon which we can perform a quantifiable analysis of the language with an objective stylometric statistical investigation. Although several texts exist for which psychics state that the language in their texts is not their own, I specifically chose Ruth Montgomery. Both Ruth Montgomery and Arthur Ford were prolific writers during their lifetime. In addition to the collaboration text, *A World Beyond*, there are accessible texts written by them individually during their lifetime. This allows for the same type of stylometric examination of language as seen in the more traditional academic literary studies of author identification such as the controversial *Federalist Papers*.

As with the *Federalist Papers*, I have a contentious block of textual language; Arthur Ford's purported language in *A World Beyond* (referred to as the unknown text). I will attempt to attribute authorship of the unknown text by comparing that block of text to Arthur Ford's living language texts, Ruth Montgomery's other texts and two additional psychic texts through a quantitative statistical analysis.

In addition to lending themselves to this type of analysis, I chose Ruth Montgomery and Arthur Ford's texts because they were written prior to the systematic editing that occurs with current writings due to increased computer technology that allows spelling and grammar checks.

I suspected that perhaps their language would be more reflective of their original writings and wanted to see if stylistic differences would occur when compared to the two more modern psychic texts of Sylvia Browne and Ramtha – JZ Knight. Both of these non-related psychics have available texts that claim to channel entities and their texts are used in this study as control texts to insure that my analyses can correctly identify author differences.

No two members of a speech community use language in precisely the same way. Therefore, an individual's idiolect contains habitual characteristics markers, which can be used to determine authorship. Combined with the advent of computers and powerful statistical software, Montgomery's claim can now lend itself to a scientific analysis by investigating the high frequency word usage of each author.

Parapsychology Expenditures

The parapsychology industry expenditures are currently unregulated and little about the profits of the industry has been published. Steven Glass in a 1988 *Harper's* article called "Prophets and Losses" claims that telecommunications analysts estimated that psychic hotlines alone grossed \$1 billion a year and those revenues were expected to double by the year 2000. The industry is extremely secretive, comprised of privately held companies that never open their books, even requiring psychics to sign non-disclosure agreements (p. 70).

In a July 19, 2005, email from Michael Shermer, Publisher and Editor-in-chief of *Skeptical* magazine, which he has graciously given me permission to quote, regarding annual parapsychology expenditures he states, "I have never found a reliable source. Typically you hear of figures in the tens of hundreds of millions of dollars but the problem is that there is no union, as such, that keeps track and it tends to be an off-the-books type of business that even the IRS

does not track”. He further suggests that the big guns of the industry such as Sylvia Browne, James Van Praagh make millions on book sales and seminars. He cites the following example. Each of them has had books on the New York Times best-seller list. A typical book contract on a hardback book is 10% for the first 5000 copies sold, 12.5% for the next 5000 copies sold, and 15% for all copies sold thereafter. That is a percent of the retail cost of the book. If Sylvia Browne sells 250,000 copies of a book that costs \$25.00 for a hardback copy, she has netted a little over \$900,000.00. Browne has many other books that continue to sell, plus her readings, workshops and seminars. Shermer guesses that at minimum the top psychics gross two to three million dollars a year (Shermer, 2005). To confirm Dr. Shermer’s point, Browne’s current web site lists 25 books for purchase, while a 25-30 minute telephone reading costs \$700.00. In addition, she has 24 seminars scheduled in the coming months with differing costs depending on the area.

While we can estimate the economic compensation of the renowned psychics, the majority of the industry (hotlines aside) is comprised of mom and pop storefront psychics. There are no statistics available as to how many of them exist or how much money is spent with them by the public. The 2005 Athens, GA telephone book listed two in-town psychics and four psychic hotline networks. The Athens psychics charge \$25.00 for a reading that could last from 20 minutes to an hour. If we project the situation in Athens across the whole country, what can reasonably be assumed about the annual parapsychology expenditures is that this is currently a multi-billion dollar a year industry.

A quantifiable scientific study that provides empirical evidence will help to address the mystique that surrounds this largely faith-based industry and either lend credibility to psychic

claims or justify the concerns skeptics express regarding the vast sums that are bilked out of general public by psychics operating in this industry.

Biography of Psychic Authors

As explained above, the psychics in this study were specifically chosen for their individual prolific lifetime writings as well as for the era in which they practiced their skill. Each of them has several best selling books from which we can document their language. In addition, each of them is well known to both lay people and colleagues involved in the parapsychology industry. In researching which psychics to study, I found that each decade appears to embrace a “psychic du jour”. Although each of them continued to practice in subsequent decades, their popularity faded as newer prophets rose to fame:

- Ford – late 1950’s – 1960’s
- Montgomery 1970’s
- Ramtha (J.Z. Knight) - 1980’s
- Browne - mid 90’s – current.

My criteria also mandated that no physical outside third party be involved in the writing of the texts. For example, probably the most renowned psychic of the 20th century was Edgar Cayce, whom I decided not to use as a candidate for this project. Unfortunately, all of the language that exists from his trances was hand recorded by stenographers and might contain contaminated data. Others such as James Van Praagh, who enjoyed immense success in the 1990’s, simply did not have enough text materials.

Ultimately, although each of the above psychics claim to channel spirit guides whose language could be compared and studied, Montgomery’s claim that a once living person (Ford)

with documented lifetime texts, makes possible my analysis. Not only can I compare the unknown language to Montgomery and the two control psychics, I can compare it to Ford's living language. Listed below are brief biographies of each subject in this study and their accomplishments:

- Arthur Ford: 1896-1971, realized his ability during WWI, emerged as a trance medium – channeling Fletcher, a dead acquaintance, who became his spirit guide. Achieved fame in 1928 by breaking the Houdini message code from the other side through his spirit guide Fletcher. Suffered from morphine and alcohol addiction for 20 years due to severe injuries sustained during an automobile accident. Again came into public prominence in 1967 during a television discussion on life after death when he went into a trance and delivered several messages that revived public interest in spiritualism. Published four books on psychic and paranormal topics. (Tribbe, 1996).
- Ruth Montgomery: 1913-2001, began her career as a prestigious Washington, DC reporter. Published an article about famed psychic Jeanne Dixon and began looking further into paranormal phenomenon. Shortly afterward found she had the gift of automatic writing by which she could communicate with various spirit guides from the other side. Published 15 books on paranormal and psychic topics (NearDeathExperiences, 2005).
- JZ Knight: born Judith Darlene Hampton March 16, 1946 in Roswell, New Mexico. Started channeling Ramtha in 1977 (Melton, 1998). Knight asserts that Ramtha is a 35,000-year-old spiritual being who was, according to Knight, "a Lemurian warrior who conquered the continent Atlantis and later became

enlightened”. Founded Ramtha’s School of Enlightenment. Numerous books have been published under the name of Ramtha and a few under JZ Knight relaying her experience as a channeling body (Wikipedia, 2005a).

- Sylvia Browne: born Sylvia Celeste Shoemaker on October 19, 1936 in Kansas, City, Missouri. Browne is a self-proclaimed psychic medium, and authoress of numerous books on spirituality. She began performing psychic readings in 1973. Browne professes the ability to speak with her spirit guides via trances that include automatic writings. She has published numerous books on paranormal and psychic topics (Wikipedia, 2005b).

Parapsychology and Language

Despite an intensive search of linguistics and parapsychology, I was unable to find any past research that links these two fields. Although numerous texts exist from psychics that claim some or all of the written language in their texts is not their own but belongs to an entity speaking through them, no linguistic studies have been made attempting to determine authorship status of such writing. What primarily exists in current literature are two opposing views about the field of parapsychology. Parapsychologists that intimate that some tests may indicate there is a sixth sense that we cannot yet explain (Radin, 2003). In contrast skeptics such as Dr. Shermer (1997) suggests in *Why People Believe Weird Things* that the person making the extraordinary claim has the burden of proving it to the experts at large and none have yet done so.

A rather distinct theory outside the realm of parapsychology involves Psychologist Julian Jaynes who proposes that psychics who experience “voices” in their minds are experiencing a residual effect of a bicameral mind, which he claims to have affected all humankind until late in

the second millennium. He maintains that psychics, like schizophrenics, are remnants of that past era and have simply not developed the current conscious awareness most people experience (Jaynes, 1976).

None of the above perspectives links language and the parapsychology, but they do serve to highlight the need for further empirical research that can determine which view might better explain the claims made by the industry.

Corpora of writings by Psychics

Eight separate corpora were built for this analysis. The breakdown of each corpus is as follows:

- *A World Beyond* – Ruth Montgomery (Montgomery, 1971). This is the text that contains the unknown language purported to be Arthur Ford, which is being analyzed for comparative purposes. The text 58,265 words. Once it was converted into Microsoft word and reviewed for accuracy, it was further broken down into two separate files. One file contains Ruth Montgomery's language and another Arthur Fords language. Arthur Ford's purported language was always delineated by quotes in the printed text. Arthur Ford's language accounted for 45,303 words of the text (77.8%) and Ruth Montgomery's language accounted for 12,962 words of the text (22.2%). In addition, Ruth Montgomery's text includes a very small portion of the supposed language of other entities (guides) (.06%) that speak to her during her automatic writing sessions. This language was attributed directly to her.

- *The World Before* – Ruth Montgomery (Montgomery, 1976).

67,534 words

- *The World to Come* – Ruth Montgomery (Montgomery, 1999).

30,740 words

- *Unknown but Known* – Arthur Ford (Ford, 1968).

49,767 words

- *Why We Survive* – Arthur Ford (Ford, 1953).

24,749 words

- *Conversations with the other side* – Sylvia Browne (Browne, 2002).

29,979 words. Sylvia Browne's language includes many instances of the contractions 'you've' and 'I'll'. These were changed in the text to reflect 'you have' and 'I will' because none of the other texts relied on contractions. Standardizing them allows for a direct comparison of word frequency with the other texts in the study. In addition, her books include first person questions to the alternate entity speaking in the book. These questions were removed from the corpus because it was unclear who wrote the questions. This is an interesting phenomenon because like Montgomery, Browne maintains that during these automatic writing sessions, she is completely unaware of the process and awakes to find the writings. I was unable to determine if these questions were formulated during the trance process or if they had been inserted during the editing process. I subsequently felt since the identity of the writer was unknown, removal would insure there was no contaminated data from an outside party.

- *Mini-Teachings* – Ramtha (J. Z. Knight) (Ramtha, 1985 - current).

65,179 words

Note: As I explained, I chose Montgomery because I suspected that perhaps *A World Beyond* might have escaped the systematic editing process provided by today's technology. Excluding Montgomery's 1999 work, Browne's work and Ramtha's Mini-readings, all of Fords and Montgomery's early writings reflected a high number of spelling errors. Proof reading indicated these were not OCR recognition errors. Editing and printing of these earlier writings prior to technology that is more recent apparently had some effect. Since this study focuses on high frequency function words and the words involving spelling inconsistencies were exclusively content words they were not changed in the individual corpora. Example: emphasises versus emphasizes, baptised versus baptized. This might be a reflection of the difference between British English and American English but both Montgomery and Ford were American and American presses published their texts.

Determining Word Frequency Hierarchy – Psychic data

Determining word frequency in any corpus is easily handled with an integrated software program that allows you to determine how words behave in texts. WordSmith Tools is such a program and it was used in this study. Although my study specifically focuses on word frequency, this program has several additional applications that can assess what language patterning may be going on within a specific corpus.

Traditional literary studies that have used the type of analysis that I apply to psychic language have determined that it is the 50 – 75 most frequent words in a text that provide characteristic evidence of authorship. John Burrows typically uses 50 – 75 word types for his

analyses. He suggests that the number of words chosen is a matter of convenience; however, the top 50 words generally make up about half of all word tokens in most written texts in English (Burrows, 1992b).

For the purposes of this study, I determined that the 50 most frequent words would comprise my hierarchy barring any anomalies based on the genre of study. Most corpora reflect a wide variety of texts that include many different genre types. The wide variety usually precludes content words from entering the hierarchal structure. This study is specific to psychic language genre and thereby I had a concern that words that normally would not make a hierarchal list in a word frequency compilation might appear. After reviewing the statistics, I examined the word frequency list of the entire corpus. Due the specificity of this particular corpus' genre, I found two words that I removed from the top most frequent word list. Neither of these words is reflective of function word usage in a non-genre corpus (Brown Corpus) and I consider them content words rather than function words and that they are not reflective of habitual characteristic evidence of a writer's individual idiolect.

‘God’ 988 tokens, 39th word on the list,
‘Life’ 759 tokens, 48th word on the list.

Once I determined the number of types, I decided that the hierarchy should be comprised of six texts excluding both Ford's language (the unknown text) and Ruth Montgomery's language in *The World Beyond*. The unknown text is the text I am investigating and I want to compare it to the texts in my hierarchy for a potential probability match, which includes Ford's living language. Also, once I reviewed the small amount of data included in the text *The World Beyond*, which reflects Montgomery's language I decided due to its minimal size and the fact I had two other texts of her living language, I would exclude that also. Her language in *The World Beyond* accounted for only 22.2% of the text and much of it was attributed to other spirit guides

she purports to channel. To determine the hierarchy I combined the remaining six texts into one large corpus.

The following chart details the statistics of the combined corpus. There are 269,759 total words, which are referred to as tokens. There are, however, only 17,150 types, which reflect actual distinct words. A distinct word such as ‘and’ is considered a type that can be used numerous times and the number of times it is used is displayed in the token total.

Table 4.1 Hierarchy Statistics – Psychic data

			Hierarchy Statistics				
N	1	2	3	4	5	6	7
Text File	Overall	WWSFord	UBKFord	WTCMont	TWBMont	CFOSBrown	Ramtha
Bytes	1,583,742	145,132	304,507	181,279	412,236	183,322	357,266
Tokens	269,795	24,799	50,198	30,829	67,938	30,312	65,719
Types	17,150	3,830	7,838	4,644	8,326	4,044	4,639
Type/Token Ratio	6.36	15.44	15.61	15.06	12.26	13.34	7.06
Standardized Type/Token	41.83	40.97	47.61	42.94	46.3	41.2	32.8
Ave. Word Length	4.47	4.4	4.66	4.48	4.69	4.59	4.13
Sentences	11,559	978	2,379	1,255	2,161	1,557	3,229
Sent, length	22.21	23.21	19.9	22.86	29.68	18.21	20.28

WWSF – Why we survive – Ford
WTC – World to Come – Montgomery
CFOS – Conversations from the other side – Browne

UBK – Unknown but known – Ford
TWB– The World to Before – Montgomery
Ramtha – JZ Knight

Although in the past word-length and sentence length have been proposed as potential characteristics of authorship, ensuing studies have met with little success. David I. Holmes (1998) in his review of the origins of stylometry cites the two seminal studies that continue to be the basis for the later studies. Mendenhall (1887) , a physicist, showed that word length generally does not reflect authorship and G. Udny Yule (1939), in his early research, concluded sentence length as a potential determiner of authorship is not entirely reliable. Extrapolating from the above chart and statistically comparing it to the unknown text corpus, which was run

separately in WordSmith Tools, we can clearly see word length and sentence length allow us to conclude little about authorship. The word length (bolded) of the unknown text falls directly between Ramtha and Ford, while sentence length (bolded) patterns to Montgomery.

Table 4.2 Word and Sentence length statistics – Psychic data

- Ford and Montgomery statistics have been averaged for their total.

LENGTH	UNKNOWN	FORD	MONTGOMERY	BROWNE	RAMTHA
Word	4.32	4.53	4.78	4.59	4.13
Sentence	25.6	21.5	26.3	18.2	20.2

There is a limitation when converting the corpus from Microsoft Word to plain text files when attempting to capture paragraph formatting in texts. In some instance, the conversion would capture paragraph formatting and in other instances, it would not. Therefore, the WordSmith paragraph statistics are not reported in the hierarchy statistics as they represent inaccurate information.

The words, their frequency (total number of times they occurred) and percentage of over all distribution within the master corpus are as follows:

Table 4.3 Hierarchy Word Frequency List – Psychic Data

HIERARCHY WORD FREQUENCY LIST							
WordSmith Tools -- 7/11/2005 12:13:29 PM							
N	Word	Freq.	%	N	Word	Freq.	%
1	THE	14,952	5.54	2	OF	8,885	3.29
3	AND	8,659	3.21	4	TO	7,480	2.77
5	THAT	5,253	1.95	6	A	5,195	1.93
7	IN	5,087	1.89	8	IS	4,442	1.65
9	YOU	4,091	1.52	10	IT	3,494	1.30
11	I	2,737	1.01	12	ARE	2,206	0.82
13	AS	2,193	0.81	14	WAS	2,161	0.80
15	FOR	2,039	0.76	16	THEY	1,872	0.69
17	THIS	1,871	0.69	18	BE	1,783	0.66
19	HAVE	1,764	0.65	20	WITH	1,684	0.62
21	WE	1,588	0.59	22	NOT	1,551	0.57
23	BUT	1,477	0.55	24	HE	1,459	0.54
25	WILL	1,441	0.53	26	ON	1,417	0.53
27	WHO	1,358	0.50	28	YOUR	1,213	0.45
29	ALL	1,208	0.45	30	SO	1,184	0.44
31	WERE	1,157	0.43	32	FROM	1,130	0.42
33	BY	1,110	0.41	34	WHAT	1,090	0.40
35	THERE	1,081	0.40	36	HAD	1,079	0.40
37	OR	1,046	0.39	38	THEIR	989	0.37
39	ONE	963	0.36	40	HIS	956	0.35
41	MY	923	0.34	42	WHEN	916	0.34
43	AN	877	0.33	44	AT	875	0.32
45	IF	842	0.31	46	WHICH	841	0.31
47	INTO	745	0.28	48	HAS	711	0.26
49	CAN	701	0.26	50	BECAUSE	692	0.26

These 50 words account for 44.3% of all the words in the master corpus. If I return the two content words I removed from the original word frequency list ‘God’ and life’, the list would account for 45.4% of all the words in the corpus – a little less than half - which falls into the realm of Burrows (1992b) guidelines regarding hierarchy.

An additional concern I had with the hierarchy is the fourth word on the list: ‘to’ 7,480 tokens. Unfortunately, limited resources do not allow me to tag the texts in this analysis for parts of speech and in the comparison phase of this analysis, I am unable to break down whether my authors use this word as an infinitive or a preposition. I felt the density of its use within the

master corpus would not allow me to remove it as I had the two earlier content words. It would clearly affect the patterning. For reference, I did want to assess an approximation of how the authors used the word and I applied following two procedures to gain an estimation of usage. First, I ran 100 lines of a concordance for the word ‘to’ in WordSmith on the master corpus. A concordance allows you to see any word or phrase in context and observe what type of company it keeps and how it is used. Once you have the concordance you can further explore the entire surrounding by growing the context for entire sentences. I broke down each usage and found that ‘to’ as an infinitive was used 52 times (52%), and as a preposition, it was used 48 (48%) times.

Table 4.4 Sample Concordance from WordSmith Tools – Psychic Data

Sample Concordance from Wordsmith		
N	Concordance	Word No.
1	nchette suddenly came to life. "Many moons a	7,511
2	2 she moved to Canada to serve as a psychiatri	13,451
3	ed withdrawal in order to help a parent, through	67,006
4	a problem that will have to be faced, for when	67,035
5	er a great deal of travel to every area of the Atla	45,664

Secondly, I ran approximately 300 words of a randomly chosen block of text from the master corpus through the UCREL Claws program (UCREL, 2005). Claws is a parts-of-speech tagger for English. It offers a free-trial service that tagged the text of 300 words. It found 17 instances of the word ‘to’, ten times as an infinitive (59%) and 7 times as a preposition (41%). Infinitives appear have a slightly higher ratio of use than prepositions within the master corpus.

Although this study does not delineate between the usages of the word ‘to’, I feel compelled to highlight this limitation to illustrate how each word in a hierarchy list must be

considered for the roll it can play in this type of analysis. The percentages listed here are based on a rough estimate of four authors in the master corpus; however, without tagging I am unable to determine how each author applies the word within their own text and how it may indeed be a characteristic marker that would be reflective of great variance amongst the authors.

Analysis Matrix – Psychic data

Having completed the hierarchal list, I built the matrix, which is used for the analysis. Each individual text was segmented into 5,000 word blocks of text. Neither an editor like Notepad nor an analysis program such as WordSmith has the capacity to segment text. I returned to the text file of each author that I had initially converted from OCR into Microsoft Word. From Microsoft Word, I used word count to cut blocks of 5000 word segments, saved them individually as a text file and subsequently read the file from WordSmith. WordSmith allowed me to search each 5,000 word segmented text file individually for the number of occurrences for each word in my hierarchy. The raw data was then entered into my matrix list. Below is an abbreviated sample of my matrix and the first four words in the hierarchy as well as the segment length when I converted the 5000 word block into a text file from the Microsoft word file. The complete matrix can be found in Appendix A.

Table 4.5 Sample matrix raw data – psychic data**Sample Matrix raw data - Psychic**

WORD RANK			1	2	3	4	
Author	Texts	Segment	THE	OF	AND	TO	Segment Length
Montgomery	1 WBM	1	241	122	137	149	5041
Montgomery	1 WBM	2	239	148	127	151	5030
Montgomery	1 WBM	3	134	88	133	78	2992
Ford	2 WBF	1	225	140	165	176	5033
Ford	2 WBF	2	218	104	140	208	5012
Ford	2 WBF	3	284	194	135	190	5023
Ford	2 WBF	4	269	177	184	155	5027
Ford	2 WBF	5	212	110	157	182	5017
Ford	2 WBF	6	207	129	173	191	5018
Ford	2 WBF	7	238	155	145	189	5034
Ford	2 WBF	8	242	153	145	177	5019
Ford	2 WBF	9	265	178	171	187	5336

The segment length is used because no complete text will divide evenly into 5000 word segments and therefore prior to analysis we need to convert the raw frequency of words in each segment to relative frequencies for accurate statistical analysis. In addition, for reasons unknown but critical to the analysis, how Microsoft Word counts words and how WordSmith counts words differs. Microsoft Word's count is a hard box and it cannot be changed. WordSmith does allow you to set some standards as to what you as a researcher you consider a word count. Therefore, standardization of the raw frequencies becomes even more important. As indicated above despite a 5000 word cut from the Microsoft Word document, WordSmith generally counted additional words. For example in Montgomery's first segment of 5000 words in the collaborative text *The World Beyond* from Microsoft word translates into 5041 words in WordSmith. Converting the segment length from raw frequency to relative frequency insures the statistical analysis is comparing apples to apples. An example of the word count difference between the two programs follows.

Table 4.6 Word count discrepancies – psychic data

Author	Texts	Segment	THE	OF	AND	TO	Segment Length
Montgomery	1 WBM	1	241	122	137	149	5041

Text File WBM1.TXT		Microsoft Word Count
Bytes	29,023	0
Tokens	5,041	5000
Types	1,495	0

Lastly, the raw data is converted using segment length to standardize the data, which reflects relative frequencies. It is this matrix that is used for multivariate analysis. Multivariate analysis generally requires that you have at minimum as many cases as you have variables. By segmenting the seven texts into 5,000 word blocks, the matrix is comprised of 50 variables (high frequency words) and 66 cases (text segments).

Table 4.7 Sample matrix relative frequencies – psychic data**Sample Matrix relative frequencies – Psychic data**

Author	Texts	Segment	THE	OF	AND	TO	THAT	Segment Length
Montgomery	1 WBM	1	0.047808	0.0242015	0.0271771	0.0295576	0.016465	5041
Montgomery	1 WBM	2	0.0475149	0.0294235	0.0252485	0.0300199	0.0151093	5030
Montgomery	1 WBM	3	0.0447861	0.0294118	0.0444519	0.0260695	0.0177139	2992
Ford	2 WBF	1	0.0447049	0.0278164	0.0327836	0.0349692	0.0176833	5033
Ford	2 WBF	2	0.0434956	0.0207502	0.027933	0.0415004	0.0211492	5012
Ford	2 WBF	3	0.0565399	0.0386223	0.0268764	0.037826	0.0167231	5023
Ford	2 WBF	4	0.053511	0.0352099	0.0366023	0.0308335	0.0185001	5027
Ford	2 WBF	5	0.0422563	0.0219255	0.0312936	0.0362767	0.0213275	5017
Ford	2 WBF	6	0.0412515	0.0257075	0.0344759	0.038063	0.0145476	5018
Ford	2 WBF	7	0.0472785	0.0307906	0.0288041	0.0375447	0.0150973	5034
Ford	2 WBF	8	0.0482168	0.0304842	0.0288902	0.035266	0.014744	5019
Ford	2 WBF	9	0.0496627	0.0333583	0.0320465	0.035045	0.0134933	5336

Analysis – Psychic Data

While each of the following methods could be their own section, I have chosen to italicize them instead. Each multivariate technique that I used (principal component analysis, discriminant analysis and cluster analysis) supplies a piece of the puzzle that ultimately completes the picture of the analysis to be explained. It is the compilation of the three techniques that allow me to draw a reasonable conclusion.

PRINCIPAL COMPONENT ANALYSIS (PCA) - My goal in using PCA analysis is to reduce a large number of variables, my fifty most frequent function words, to a more manageable number of components (factors) from which I can make inferences about my data as well as utilize them in other statistical techniques such as discriminant analysis. Craig and Burrows in their 2001 article *Lucy Hutchinson's Authorship: A Computational Approach* suggest that for a PCA analysis, the first two principal components are generally sufficient to reflect the most interesting variables to observe and determine which correlate more highly with one factor or another (p. 263). Additional components will be included in discriminant analysis to determine authorship; however, PCA will only need to use component 1 and 2 to investigate how the function words behave and which ones have become characteristic markers for individual authors.

The PCA analysis on the Psychic Language data indicated that the first two factors of the total variance could explain 34.2% of the total variance within the data. The first principal component accounted for 23.9% of the variance and the second factor 10.3%. Listed below is the Total Variance Explained chart from SPSS.

Table 4.8 Total Variance explained – principal components 1 and 2 – Psychic data

Total Variance Explained – Principal Components 1 and 2						
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	11.929	23.857	23.857	11.929	23.857	23.857
2	5.167	10.334	34.191	5.167	10.334	34.191

Extraction Method: Principal Component Analysis.

Having determined the two components (factors), I attempted to examine the underlying relationships between the variables and draw inferences about how the authors use the variables. The beauty of PCA is that no mathematical assumptions are needed as the variables are allowed to choose themselves. By taking a sampling of authors texts (Montgomery, Ford, Brown and Ramtha), I am allowing the unknown text to find its place within the genre of Psychic language based on the frequency of function words. Once PCA is applied, the first two components comprise the x-axis and y-axis in figure 1a. Figure 4.1a allows us to see how the function words behave. Variables that appear together tend to behave alike. In addition, they tend to be found more often in one group of texts and conversely are found less often in the other groups of texts. Burrows and Craig (2001) also suggest that the variables that are found at either end of the end of a specific component tend to be the most important in forming the two components. In this specific analysis, the variables on the x-axis ‘because’ ‘you’, ‘it’, ‘what’ are the words that are found to the far right end of component 1 of the plot in figure 4.1a and ‘with’, ‘by’ toward the left end of the plot suggesting they were most important in forming the first component. Conversely, the words ‘we’, ‘were’ ‘was’ ‘or’ correlated most highly with component 2 and can be found at the top and bottom of the y-axis.

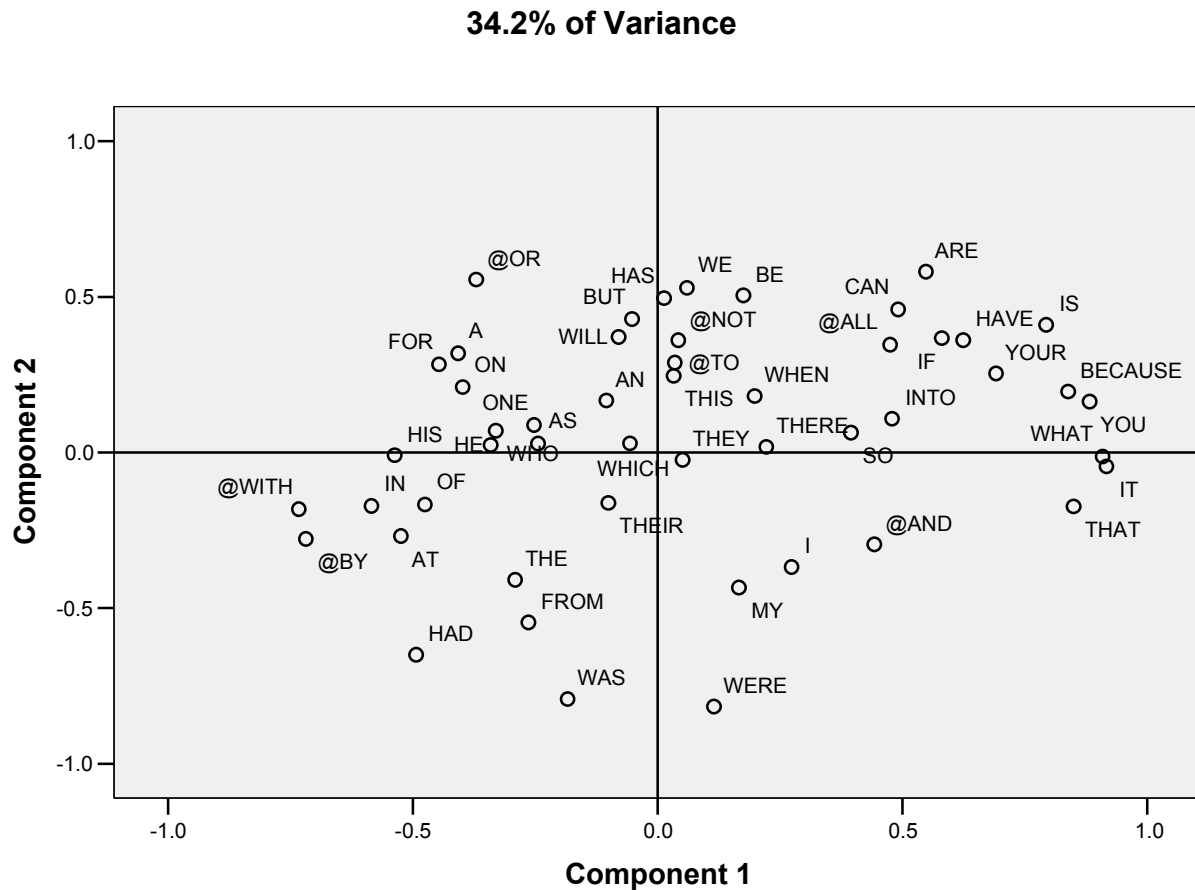


Figure 4.1a – Psychic data word plot

Having plotted how the 50 most frequent function words behave, further analysis is needed to ascertain how much of each component (1 and 2) is found in the individual segments (66 – 5000 word text entries) that represent the four authors and the unknown text. Figure 4.1b shows where each author's individual segments fall within the word plot and provides a statistical group centroid that calculates where the compilation of an individual author's segments (text entries) fall in the plot.

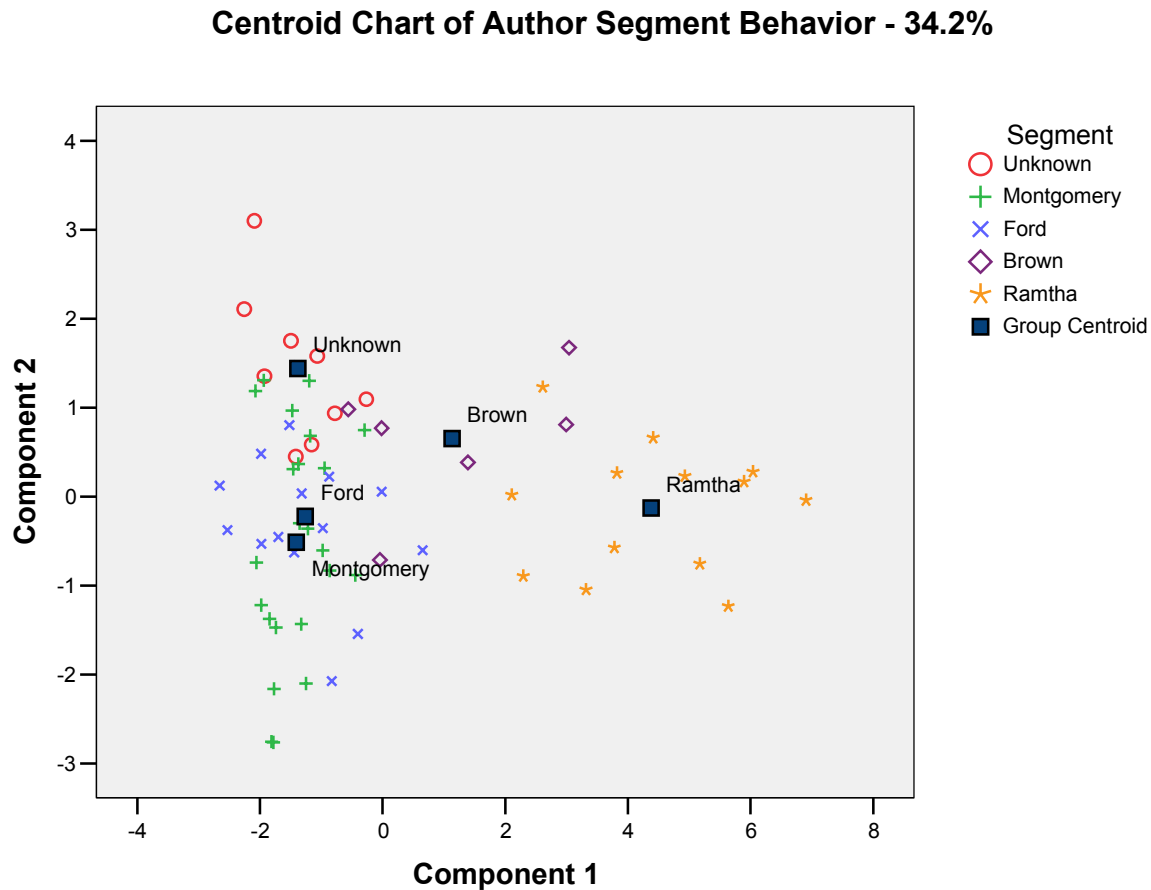


Figure 4.1b – Psychic author segment behavior

Figure 4.1b plots the text entries by author segment and compiles them into a group centroid on the same two components found in figure 4.1a. In other words, figure 4.1a represents a statistical picture of how the 50 function words behave, while figure 4.1b represents a statistical picture of how the actual author's texts behave based on 34.2% of the data variance.

The horizontal axis in Figure 4.1b confirms that the Psychic Language study involves five distinct authors and that they clearly use the 50 function words differently. Based on the placement of the group centroids, we find Ford, Montgomery and the unknown author on the far

west side of the plot while Brown sits in the middle and Ramtha far to the east. It is apparent that Ford, Montgomery and the unknown author's language use differs most greatly from Ramtha and distinguishes itself to a lesser degree, from Brown.

The vertical axis in figure 4.1b indicates that the Ford, Montgomery and the unknown text reflect some function word usage similarities but still with notable stylistic differences. The initial picture painted in figure 4.1b would suggest that Ford's use of language might be closer to the language found in the unknown text; however, caution must prevail. The above plot only allows us to view 34.2% of the variance within the data. In addition, the individual segments, outlined in figure 4.1b, indicate that Montgomery has segmented texts that fall over a much wider range on the vertical axis than Ford and the unknown author's texts. Montgomery's work intersperses itself among both of the other authors and at this juncture, no precise conclusion can be drawn as to whose voice is found in *The World Beyond* without further multivariate analyses.

Although no concrete conclusion can be drawn regarding authorship, before moving on to additional multivariate analyses, PCA does provide an avenue to investigate the underlying way the 50 function words behaved. This process details which of the 50 words correlated most highly with component 1 compared to which words correlated mostly highly with component 2. From a qualitative perspective, the procedure provides insight as to what specific words were involved in creating the variance within the authors in the two components and suggests investigation to determine if there are any identifiable patterns that could explain what characteristic markers best explain idiolect. Table 4.9 is a matrix of words that correlated most highly with component 1 and 2.

Table 4.9 Rotated Component Matrix – Psychic Data**Rotated Component Matrix – Psychic Data**

	Component	
	1	2
IT	.917	
WHAT	.909	
YOU	.883	
THAT	.850	
BECAUSE	.839	
IS	.794	.410
@WITH	-.733	
@BY	-.719	
YOUR	.691	
HAVE	.624	
IN	-.585	
IF	.581	
HIS	-.537	
AT	-.524	
CAN	.491	.460
INTO	.479	
OF	-.476	
@ALL	.475	
FOR	-.447	
WERE		-.817
WAS		-.792
HAD	-.494	-.650
@OR		.556
FROM		-.546
WE		.529
BE		.505
HAS		.496
MY		-.434
BUT		.429
THE		-.409

Extraction Method: Principal Component Analysis.

Components that are defined by just one or two variables provide very little information and a common rule of thumb is that each component should have at least three variables that correlate highly in the component for interpretation. Both negative and positive correlations have equal investigative value, as it is the absolute value that is considered. The closer the

number is to 1, the more it contributed to the variance of the component. In the above table, the word 'it' correlated the highest to component 1 while the word 'were' correlated highest to component 2.

An analysis of the speech class each word belongs to in Table 4.9 does not allow us to assign meaning to the component as a whole. The classes do not hold any identifiable patterns with the exception that prepositions appear to comprise the highest majority of words in component 1, accounting for 37% of the speech class variance, while mostly past tense verbs impact component 2 comprising 42% of the speech class variance. In addition, interestingly subordinate conjunctions appear to account for the variance in component 1, while coordinating conjunctions affect variance in component 2.

TABLE 4.10 Principal Component 1 Parts of Speech – Psychic Data

Principal Component 1 – Parts of Speech Psychic Data

Modal	Pronoun	Noun	Determiner	Verb	Sub Conj.	Preposition
Can	You	It	What	Is	Because	With
	His		That	Have	If	By
			Your			In
			All			At
						Into
						For
						Of

Principle Component 2 – Parts of Speech Psychic Data

Article	Pronoun	Determiner	Verb	Coord Conj.	Preposition
The	We	My	Were	But	From
			Was	Or	
			Had	And	
			Be		
			Has		

As we look at the list, it is apparent that many of the component 1 variables are found to the extreme right in figure 1a on the horizontal axis and they can be attributed to the author Ramtha and to a lesser degree Brown in figure 4.1b. Conversely, variables are found to the left of the figure 1a on the horizontal axis are reflective of Ford, Montgomery and the unknown text in figure 4.1b. Unfortunately, it is more difficult to assign word usage when looking at the vertical axis in component 2. To determine why the authors word choices broke down in this manner would require an intense literary study of their individual narrative styles and is outside the realm of this study. The chart, however, does allow us to glimpse what words in this study became characteristic markers for the individual authors and confirms that individual frequencies of function word usage reflect ingrained linguistic habits.

DISCRIMINANT ANALYSIS (DA) – My goal in applying DA is to classify the segmented texts of my authors into uniquely defined populations and predict authorship of the unknown text. To apply DA, I will use the components (factors) that were developed in PCA, which reduces the large number of variables (function words) to a more manageable number of components that contain the variance within the data. Because PCA calculates as many principal components as there are variables, it is necessary to decide how many components (factors) are required to adequately represent the data in DA. There would be no gain to the methodology by replacing all 50 variables in PCA with 50 components. Remember the goal is to explain as much of the variance as possible using as few components as possible. A criterion that is commonly used is the eigenvalue-greater-one criterion, which suggests that only components with a variance of 1 should be included.² Norušis explains that components with a total variance of less

² Eigenvalues - a statistic that quantifies variation in a group of variables and its accountability by a particular factor. An eigenvalue is the sum of squared values in the column of a factor matrix (www.siu.edu/~pohlmann/factglos).

than 1 are no better than using the individual variables, which do not serve to explain the overall variance of the data (2005). The total column under Initial Eigenvalues in Table 4.11 suggests that the first twelve components are adequate for this analysis.

The twelve components account for 79.5% of the total variance in the data. The remaining 38 components (20.5%) are so small that they do not contribute significantly to explaining variance. Table 4.11 is the total variance explained chart from SPSS that details the twelve new components (variables) that will be used in the DA. . The unknown text is not used in the analysis of Table 4.11 because the DA will predict to which author the nine unknown texts segments will gravitate.

Table 4.11 Total Variance Explained – Principal Components 1-12, Psychic Data

Total Variance Explained – Principal Components 1-12, Psychic Data

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	11.929	23.857	23.857	11.929	23.857	23.857
2	5.167	10.334	34.191	5.167	10.334	34.191
3	4.446	8.892	43.083	4.446	8.892	43.083
4	4.122	8.244	51.327	4.122	8.244	51.327
5	3.050	6.099	57.426	3.050	6.099	57.426
6	2.475	4.951	62.377	2.475	4.951	62.377
7	1.895	3.789	66.166	1.895	3.789	66.166
8	1.651	3.302	69.468	1.651	3.302	69.468
9	1.515	3.029	72.498	1.515	3.029	72.498
10	1.325	2.649	75.147	1.325	2.649	75.147
11	1.173	2.345	77.492	1.173	2.345	77.492
12	1.022	2.044	79.536	1.022	2.044	79.536
13	.939	1.878	81.414			
14	.910	1.820	83.234			
15	.865	1.730	84.964			
16	.773	1.546	86.509			
17	.644	1.287	87.797			
18	.611	1.221	89.018			
19	.551	1.103	90.121			
20	.507	1.013	91.134			
21	.431	.862	91.996			
22	.404	.808	92.803			
23	.369	.738	93.541			
24	.344	.689	94.230			
25	.300	.601	94.831			
26	.292	.584	95.415			
27	.275	.551	95.966			
28	.245	.489	96.455			
29	.241	.482	96.937			
30	.180	.360	97.297			

31	.167	.335	97.632			
32	.153	.306	97.938			
33	.141	.282	98.220			
34	.128	.257	98.476			
35	.119	.237	98.713			
36	.100	.199	98.913			
37	.095	.190	99.102			
38	.077	.154	99.257			
39	.071	.143	99.399			
40	.067	.134	99.533			
41	.057	.115	99.648			
42	.047	.094	99.742			
43	.037	.074	99.816			
44	.025	.051	99.867			
45	.017	.034	99.901			
46	.013	.027	99.928			
47	.013	.025	99.953			
48	.011	.022	99.976			
49	.008	.015	99.991			
50	.004	.009	100.000			

Extraction Method: Principal Component Analysis.

Prior to running the analysis, we must determine whether our predictors (50 function words) indicate significant differences in the means among our four authors. In other words, we want to determine whether our function word variables are good predictors of author variance. The significance column (bolded) should reflect a p value $< .05$. As table 4.12 illustrates, with the exception of six variables, which have been bolded, the statistics indicate a significant difference in the means among the predictors used in the analysis. The words ‘I’, ‘be’, ‘we’ ‘but’, ‘my’ and ‘when’ do not appear to differ significantly among the four authors

Table 4.12 Tests of Equality of Group Means – Psychic Data

Tests of Equality of Group Means – Psychic Data

	Wilks' Lambda	F	df1	df2	Sig.
THE	.776	5.108	3	53	.004
OF	.670	8.716	3	53	.000
@AND	.505	17.318	3	53	.000
@TO	.727	6.625	3	53	.001
THAT	.201	70.108	3	53	.000
A	.572	13.221	3	53	.000
IN	.610	11.298	3	53	.000
IS	.320	37.470	3	53	.000
YOU	.233	58.313	3	53	.000
IT	.209	66.797	3	53	.000

I	.913	1.678	3	53	.183
ARE	.648	9.609	3	53	.000
AS	.843	3.293	3	53	.027
WAS	.711	7.196	3	53	.000
FOR	.417	24.657	3	53	.000
THEY	.811	4.104	3	53	.011
THIS	.645	9.707	3	53	.000
BE	.980	.360	3	53	.782
HAVE	.609	11.341	3	53	.000
@WITH	.502	17.553	3	53	.000
WE	.969	.559	3	53	.644
@NOT	.780	4.974	3	53	.004
BUT	.944	1.058	3	53	.375
HE	.802	4.372	3	53	.008
WILL	.902	1.919	3	53	.138
ON	.622	10.738	3	53	.000
WHO	.814	4.024	3	53	.012
YOUR	.299	41.485	3	53	.000
@ALL	.759	5.597	3	53	.002
SO	.596	11.989	3	53	.000
WERE	.729	6.568	3	53	.001
FROM	.827	3.685	3	53	.017
@BY	.462	20.557	3	53	.000
WHAT	.213	65.253	3	53	.000
THERE	.836	3.469	3	53	.022
HAD	.597	11.902	3	53	.000
@OR	.474	19.574	3	53	.000
THEIR	.790	4.689	3	53	.006
ONE	.678	8.381	3	53	.000
HIS	.632	10.300	3	53	.000
MY	.947	.998	3	53	.401
WHEN	.876	2.503	3	53	.069
AN	.613	11.153	3	53	.000
AT	.668	8.772	3	53	.000
IF	.605	11.525	3	53	.000
WHICH	.837	3.440	3	53	.023
INTO	.734	6.398	3	53	.001
HAS	.619	10.891	3	53	.000
CAN	.465	20.307	3	53	.000
BECAUSE	.345	33.518	3	53	.000

In addition to testing the viability of the predictors (variables), we want to confirm whether there are significant differences among the authors across the functions words. In other words, are the four authors using the variables differently? The Wilks' Lambda test predicts whether the predictors (variables) differentiated significantly among the authors. This test is significant at the $p < .05$ and Table 4.13 indicates there are significant differences among the authors across a compilation of all predictor variables in two functions that DA uses.

Table 4.13 Wilks' Lambda – Psychic Data**Wilks' Lambda – Psychic Data**

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.001	315.377	36	.000
2	.024	179.200	22	.000

Having confirmed that the variables and the authors have significant differences, I proceeded to the classification results of the individual authors segments of texts and the nine segments of unknown text. Table 4.14 indicates that the original group cases (segmented texts of each of the four authors) classified correctly. For the nine unknown text segments (ungrouped cases), SPSS has computed discriminant scores and predicted group membership based on the original four authors.

Table 4.14 Classification Results – Psychic data

Classification Results (a)		PREDICTED GROUP MEMBERSHIP					
Original	Count	SEGMENT	Montgomery	Ford	Brown	Ramtha	Total
		Montgomery	23	0	0	0	23
		Ford	0	15	0	0	15
		Brown	0	0	6	0	6
		Ramtha	0	0	0	13	13
		Ungrouped cases	8	0	0	1	9
	%	Montgomery	100.0	.0	.0	.0	100.0
		Ford	.0	100.0	.0	.0	100.0
		Brown	.0	.0	100.0	.0	100.0
		Ramtha	.0	.0	.0	100.0	100.0
		Ungrouped cases	88.9	.0	.0	11.1	100.0

a. 100.0% of original grouped cases correctly classified

My DA analysis of the language indicates that of the nine unknown text segments in question, there is an 88.9% probability that eight of the segments reflect the language of Ruth Montgomery. There is an 11.1% probability that the remaining segment in question reflects

Ramtha's language. Most significantly there is a 0% probability that any of the nine unknown segments reflect the language of Arthur Ford as claimed by Ruth Montgomery in *The World Beyond*.

Figure 4.1c visually illustrates how tightly each authors various segments respectively group and how the nine segments of the unknown text behave in the environment of this analysis.

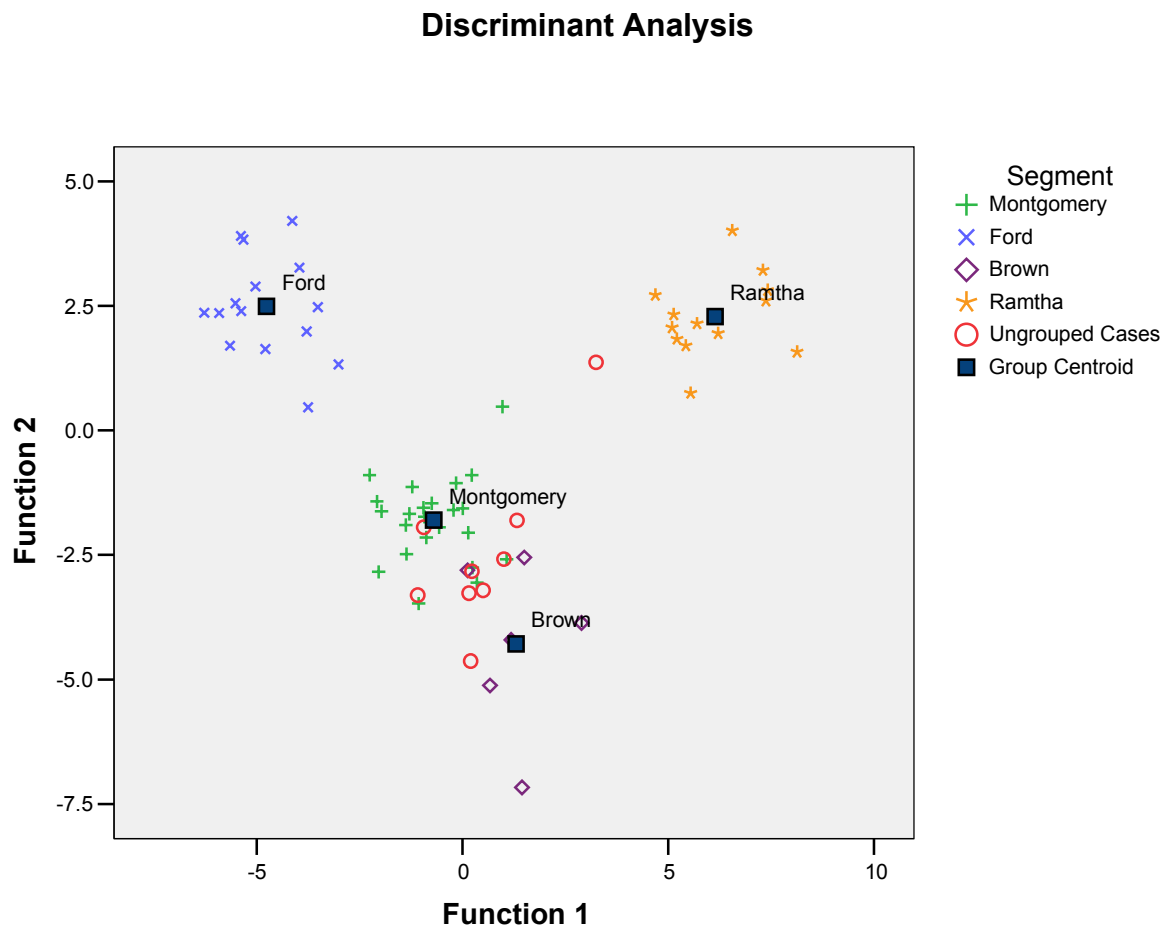


Figure 4.1c – Psychic Discriminant Analysis

CLUSTER ANALYSIS (CA) – I used DA to classify the nine segments of the unknown texts into predicted group membership. My goal in using CA is that I do not have to make any assumptions about the underlying data. I will allow the statistics to form groups based on the characteristics of the individual segments of each author including the nine segments of the unknown texts. In essence, I am operating under the assumption that I do not know from which subgroup (individual author's segments including the unknown segments) originate. This methodology will potentially confirm my findings in the DA analysis and supplement my conclusions as to whose language is found in *The World Beyond*.

I used Hierarchical Clustering, which is one of the most straightforward methods in CA. It begins with each case (segment) being a cluster unto itself and at each successive step, similar clusters are merged. A graphic representation of the distance at which clusters combine is called a dendrogram. Dendrograms are read from left to right. Figure 4.1d represents the findings of my CA analysis on the four authors segments and the unknown text segments.

What is immediately apparent is that each of the four author's segments cluster with each other indicating the uniqueness of their individual language in the usage of the function words in this study. Their individual segments consistently find each other in the first vertical line that corresponds to the smallest rescaled distance as well as in the second vertical line as similar clusters merge. For example, Montgomery's segments 21, 24, 25, 20, 22, 26 and 19 comprise the first vertical line on the left and represent similarity. Montgomery's segments 26 and 27 represent the second cluster on the first vertical line on the left. The next cluster 23, also belonging to Montgomery, identifies itself most strongly on the second vertical line on the left with her first cluster of segments 21, 24, 20, 22, 26 and 19.

The nine unknown text segments are dispersed throughout the dendrogram. They closely mirror what we found in the Discriminant Analysis of the same data.

The unknown text segment 9, clusters most strongly with Montgomery's text segment 32 and subsequently clusters with her text segments 31, 14, 16 and 15.

The unknown text segments 6, 7, 2, cluster with each other then identify with the unknown text segment 5. The unknown text segments 1 and 8 cluster with each other and subsequently cluster with Montgomery's segments 29, 30, 17 and 18. Montgomery's segment 12 successively ties together all of these segments as one cluster.

The unknown text segment 4 clusters first with Brown's segments 51, 52, 50 and 53, while the unknown text segment 3 appears to be an outlier clustering with little similarity to any one author.

Without any underlying assumptions about the data, cluster analysis confirms the discriminant analysis findings. Seven of the nine unknown text segments identify most closely with Ruth Montgomery's language, one segment clusters with Brown and one segment is open to interpretation but it is most similar to Montgomery, Brown or Ramtha. Most significantly, again none of the unknown text segments identifies themselves with Arthur Ford's language. Arthur Ford's language continues to isolate itself from the nine unknown texts purported to be his language by Ruth Montgomery in the book *The World Beyond*. This reaffirms what the discriminant analysis indicated when we made the assumption that we knew what segments belonged to what authors and we wanted to find out how and where the nine unknown segments would fall based on their function word usage.

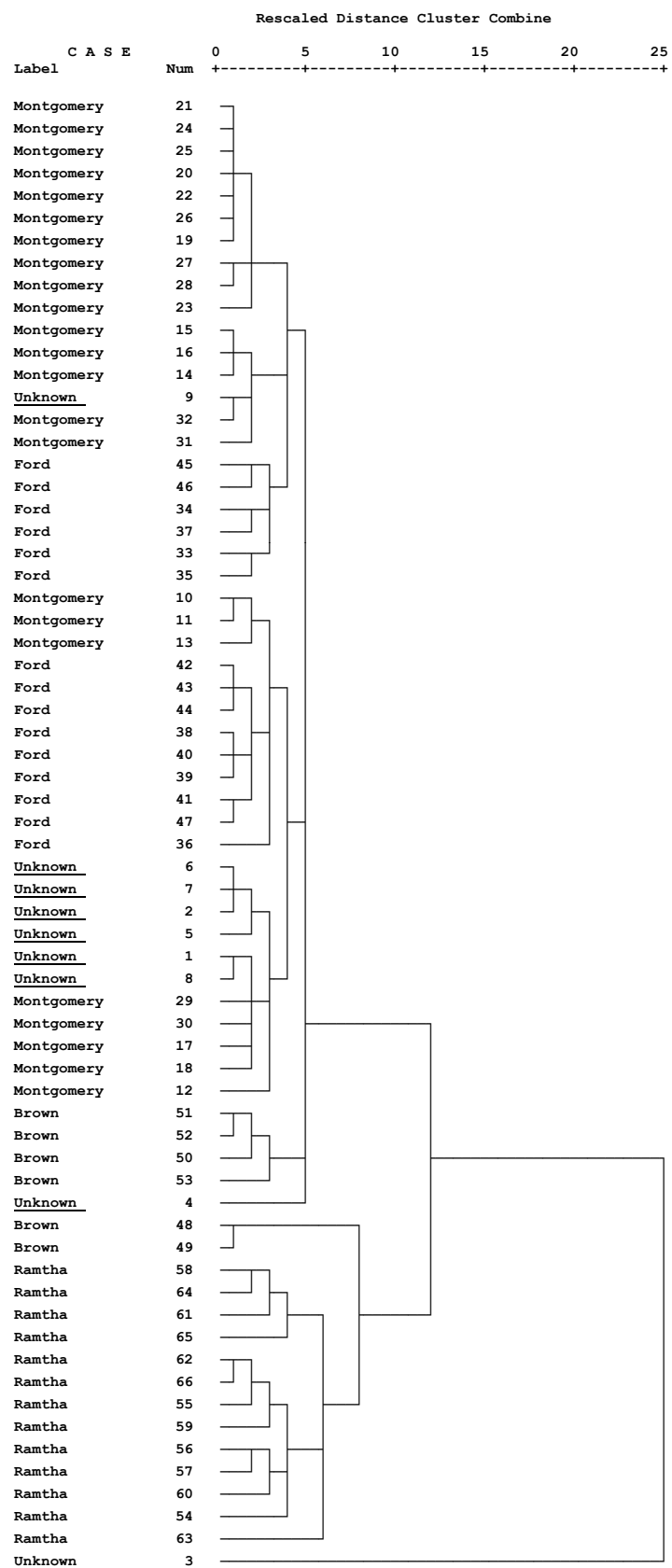


Figure 4.1d Psychic Dendrogram – cluster analysis

My quantitative analysis of the language found in *The Word Beyond* indicates that there is minimal probability that Arthur Ford's spirit wrote the unknown text, despite Ruth Montgomery's claims. Both the DA and CA analyses suggest that he is not a potential match for the nine unknown segments of text. Further, there is an 88.9% probability that the language in *The World Beyond* is none other than Ruth Montgomery's own. Ruth Montgomery's individual idiolect contains habitual characteristic markers that simply cannot be concealed when the unknown text segments are allowed to find their place within the language in this study based on the frequency of function word usage.

As a researcher, I began this study with the anticipation that my findings would reflect that the characteristic markers found in the channeled Ford text segments would echo those found in Montgomery's writings rather than match those of his original life writings. As the product of a culturally rich upbringing, abounding with tales of gypsy curses, late night ancestral visits, and annual tea leaf readings, however, I confess to harboring a small hope that perhaps Ford's language would prevail. Clearly, this case study determines to a high probability whose voice is found in *The World Beyond*, and effectively illustrates how individual function word usage can be empirically investigated with significant results regarding authorship. Linguistics is the scientific study of language and the scientific evidence in this study, which utilizes the multivariate triad, suggests that the concerns skeptics express regarding the vast sums of money that are bilked out of the public by the parapsychology industry might well be justified.

CHAPTER 5

GENRE STUDY

Fiction versus non-fiction

On January 22, 1996, a book called *Primary Colors* sent a firestorm through the world of United States politics. Published under the pseudonym Anonymous, it was an unflattering *roman à clef* about President Bill Clinton's 1992 presidential campaign. With an initial printing of only 62,000 copies, it generated such tremendous interest it spent nine weeks as number one on *The New York Times* bestseller list. Speculation raged in Washington DC as to who Anonymous might be and political pundits were convinced that Anonymous was a political insider.

Within two weeks on February 2, 1996, *The Washington Post* published "Wanted Anonymous: Sure, They Deny It, But If They Didn't Do It, Who Did?" which was a clip and save odds guide of the top thirty-five primary suspects in the author whodunit. Halfway down on the guide, Joe Klein, a Newsweek columnist and CBS consultant/commentator, with 50-1 odds gave a ritual denial stating, "I am Spartacus. All of who are accused of this should stand up and say, 'I am Spartacus.' And share in the royalties" (p. B01).

As speculation rose to a fever pitch, Don Foster, a Vassar College English professor known for his computer identification of William Shakespeare as the author of an obscure 400-year-old poem, claimed that after applying the same type of computer analysis to *Primary Colors*, the analysis showed the author to be political columnist/commentator Joe Klein. On

February 16, 1996, *The New York Daily News* reported that Foster maintained he ran millions of words written by the top suspects through a computer and that Klein was the only possible culprit. Foster singled out repeated use by both Klein and Anonymous of such words as ‘fella’, ‘mush’, ‘scruffy’, ‘lugubrious’, parallel constructs of quirky sentences, colon and dash use, and an extra letter interjection in expressions such as ‘ahh’ and ‘aww’ as keys to Klein’s identity. Joe Klein promptly denied the findings even going so far putting a message on his office answering service “to all of you who are calling about *Primary Colors*” - denied he wrote the book and then added as a salutation “good hunting.” Conjecture, rumor and gossip continued haunting some of President Clintons closest advisors including Mandy Gunwald and George Stephanopoulos as well as top Washington reporters such as Maureen Dowd, Michael Kelly and even cartoonist Garry Trudeau (Kennedy & Kornblut, p. 2).

Then in mid-July of 1996, *The Washington Post* discovered Joe Klein’s handwriting on a manuscript copy of the novel. On July 17, 1996, one of the biggest political and publishing mysteries ended when in a hastily called press conference Joe Klein without apology stated, “I am Joe Klein and I wrote *Primary Colors* (Carvajal, p. 23). Joe Klein subsequently resigned from his consultant/commentating position at CBS, took a long absence from his Newsweek column as the speculation frenzy ended and he found himself instead the subject of media ethics and journalistic credibility debates within the political beltway.

Problem Statement

Unlike the previous study on psychic data, there is no question of authorship in reference to *Primary Colors*. Joe Klein is the author Anonymous. What is in question is if the quantitative statistical methodology used in the Psychic study can successfully cross genre lines and

accurately predict authorship of *Primary Colors*. Douglas Biber in *Dimensions of Register Variation* explains that it is well known that not only differences between authors but also differences in register or text type are reflected in the relative frequencies of linguistic variation (1995). In considering questions of authorship, we need to have some idea of how genre will affect a quantitative study where the potential suspect's texts are written in different registers as in the case of Joe Klein's non-fiction political columns. These findings will serve to help identify what linguistics can quantifiably determine regarding future authorship concerns that entail genre issues.

Don Foster's computer analysis indicates that the genre of the writing does not influence its ability to determine authorship. Unfortunately, Don Foster does not share any replicable methodology with his readers, either in his article in *New York Magazine* or in his book *Author Unknown* (Foster, 2000, p. 53-94). It appears from his statements that unlike the formal quantitative methodology I am applying, the basis for his findings relies on the more qualitative stylistics of the writing, i.e. spelling, punctuation, syntax, characteristic phrasing.

Genre is a critical issue in determining authorship. As Baayan, Tweedie and Van Halteren suggested "...for one author differences in register can be much stronger than differences within a register between texts of different authors" (1996, p. 122) In other words, texts in different genres (political writings versus fiction) by one author might differ more than texts by different authors in the same genre. I contend that function words are thought to be outside the consciousness of authors and that function word usage thereby reflects deeply ingrained linguistic habit. In essence, function words are reflective of an author's uncontrollable word print despite the genre in which an author chooses to write. In the case of Joe Klein, a quantitative statistical study of his political written language versus his fiction written language

will reveal him as the author of *Primary Colors* when compared to an alternate political suspect at the time, columnist/reporter Michael Kelly.

Background of the Study

Although several texts exist upon which authors write in various genres, I specifically choose Joe Klein and *Primary Colors* for its notoriety and for the voluminous amount of text available in his political non-fiction genre.

For the comparative purposes of the study, I choose columnist/reporter Michael Kelly. Michael Kelly ranked high on the suspect list in the 1996 speculation frenzy of the author whodunit. Since authorship is no longer in question, I felt that since he was also a prolific political columnist at the time, he matched in genre to Klein. Also, he had available text as a *New York Times* reporter around the same period, and that added a third aspect to the challenge of a genre study: political columns, fiction, and news reporter. Lastly, Don Foster successfully eliminated Michael Kelly as Anonymous, after he considered him a serious contender in his qualitative investigation (2000, p. 64) and I wanted to determine how close he would rank as the author in a quantitative investigation.

Pseudonymic Authorship

Maira Allen in “Should you use a Pseudonym” explains that there are many reasons why authors use pseudonyms, from legal reasons to simple reasons like your name does not lend itself to a genre studies expectation – “Barbie Doll” as the name of an author in a scientific journal. Dean Koontz, for instance, uses several pseudonyms because he believes it allows him to change his voice and actual writing style (2004). Steven King (2004) states he chose to write under a

pseudonym in his early career because publishing companies once believed that audiences would only accept one book a year from authors. Anne Rice (2004) used a pseudonym when she wanted to write a type of erotica she herself wanted to read.

Reporter Caryn James of *The New York Times* wrote that Joe Klein stated that he wrote under the pseudonym Anonymous because he wanted the book to be judged on its merits. She points out, however, that he did not mention that it was a clever and lucrative marketing ploy that enabled him to continue writing journalistically about the Clintons without losing his sources (1996). Whatever Joe Klein's reasons, they provide us with a perfect forum upon which to analyze how language will react in cross-genre situations, and what we can expect as linguistics in forensic investigative studies.

Biography of Genre Authors

Listed below are brief biographies of Joe Klein and Michael Kelly. Unfortunately, Michael Kelly died on April 3, 2003, while on assignment in Iraq, the first American reporter killed during the conflict.

- **Joe Klein** (b. 1947) - provided by Time Magazine. Joe Klein joined TIME magazine in January 2003 to write a regular column on national and international affairs. His column, titled "In the Arena," appears in Time's upfront "Notebook" section. Klein is a senior writer based in New York and Washington, D.C. As "Anonymous," Klein wrote the critically acclaimed novel *Primary Colors*, a best-seller inspired by the 1992 political race. He has written articles and book reviews for *The New Republic*, *The New York Times*, *The Washington Post*, *LIFE*, *Rolling*

Stone and other publications. Klein wrote a column called "Public Lives" for *Newsweek* in the early 1990s; served as a consultant for CBS News providing commentary (1992-1996). Klein graduated from The University of Pennsylvania with a degree in American civilization. He lives with his wife and two children in Westchester County, New York, and is also the father of two adult sons. (TimeMagazine, 2006).

- **Michael Kelly** (1957–2003) – provided by Atlantic Media (Michael Kelly Media award). Michael Kelly was born in Washington, D.C. in 1957. After graduating with a degree in history, he worked for years as a researcher, booker and associate producer. From 1983 – 1992 he was a political reporter for various magazines and newspapers including *The New York Times*. In 1996, he became the editor of *The New Republic* and wrote the TRB column for the magazine. He joined the *National Journal* in 1997 and also wrote a weekly column for *The Washington Post*. In 2002, he became the editor-at-large of *The Atlantic*. Michael Kelly died on April 3, 2003, while on assignment in Iraq, the first American reporter killed during the conflict. He is survived by his wife Madelyn and two young sons (AtlanticMedia, 2004).

Corpora - Genre

I built four separate corpora for this analysis. The novel *Primary Colors* was built using the OCR methodology discussed in the Psychic Language Study. The texts used for Joe Klein and Michael Kelly were compiled from electronic sources that had the available political

columns and newspaper articles in their data banks. The breakdown of each corpus is as follows and the specific columns and newspaper articles can be found in the reference list:

- *Primary Colors* – Anonymous (Anonymous, 1996).
137,943 words
- *Public Lives* – Joe Klein (Klein, 1992-1996)
151 columns – cited in Reference List.
142,581 words
- *The New Republic Column* – Michael Kelly (Kelly, 1995-1997)
51 columns – cited in reference list.
70,190 words
- *The New York Times* – Michael Kelly (Kelly, 1991-1995).
32 *New York Times* articles – cited in reference list
65,998 words

Prior to determining the word frequency hierarchy of the function words to be analyzed, I did modify each corpus by removing common contractions. Although contractions can provide some distinctions in text types across genre, we know that we are comparing two dimensions – fiction, a novel that is heavily populated with conversation and relies on contractions versus non-fiction political writing, which does not. This study attempts to determine authorship across the genre. Standardized data that reflects actual function word usage levels the analysis field. In his 2003 article “Questions of Authorship: Attribution and Beyond” J. F. Burrows, whose multivariate research centers on literary studies, suggests, “with texts of a bygone era, it is usual and desirable to... expand contracted forms of expression in order to reduce the influence of trivial or accidental variations” (p. 5 - 32). I feel this also applies to a genre study where parody

of function word usage is critical. I expanded the following common contractions in each individual corpus using the guideline found on John's ESL community and interactive website for ESL teachers and students (2006):

Table 5.1 Expansion table of common contracted forms

Non- Contracted Form	Contracted Form	None Contracted Form	Contracted Form
Are not	Aren't	You have	You've
Cannot / can not	Can't	You will	You'll
Could not	Couldn't	You would	You'd
Did not	Didn't	He is	He's
Does not	Doesn't	He will	He'll
Do not	Don't	He would	He'd
Has not	Hasn't	She is	She's
Have not	Hasn't	She will	She'll
Had not	Hadn't	She would	She'd
Is not	Isn't	It is	It's
Should not	Shouldn't	It will	It'll
Were not	Weren't	We are	We're
Will not	Won't	We have	We've
Would not	Wouldn't	We will	We'll
I am	I'm	We would	We'd
I have	I've	They have	They've
I will	I'll	They will	They'll
I would	I'd	They would	They'd
They are	They're	You are	You're

Determining Word Frequency Hierarchy - Genre

Using WordSmith Tools, I determined the word frequency hierarchy that would comprise the function words in the genre study. The author Anonymous is whom I am trying to identify, therefore *Primary Colors* is not used in the hierarchy determination. Klein's column "Public Lives" in *Newsweek*, Kelly's column in *The New Republic* and Kelly's reporting in *The New York Times* comprise the corpus for determining hierarchy (referred to as hierarchy texts).

As discussed, traditional studies suggest that the top 50 words of a corpus are sufficient for a multivariate analysis. The top 50 function words within the hierarchy texts accounted for 40.3% of the word tokens in the corpus. As this entire study is politically genre based, I was not concerned with anomalies within the texts that would require removal of any tokens with the exception of names. All of the text segments within the hierarchal corpus are politically based and they were written during the 1990's. Then only word type that appeared that would skew the data when compared to the fiction novel was President Clinton's last name – 'Clinton'. It appeared as the 38th word on the hierarchy texts list and was removed to accommodate the 51st word 'so'.

The following chart details the statistics of the combined genre corpus. There are 284,203 total words (tokens) and 19,901 actual distinct words (types).

Table 5.2 Hierarchy statistics – Genre data

Text File	OVERALL	PUBLIC~1.TXT	MKELLY~1.TXT
Bytes	1,691,018	880,392	810,626
Tokens	284203	146067	138136
Types	19901	13895	13150
Type/Token Ratio	7	9.51	9.52
Standardized Type/Token	55.24	57.11	53.26
Ave. Word Length	4.72	4.79	4.69

PUBLIC~1 = Joe Klein, "Public Lives", *Newsweek*.
York Times

MKELLY~1 = *New Republic/New*

As discussed in the Psychic study, although word-length and sentence length have been proposed as potential characteristics of authorship, ensuing studies have met with little success. Word-length tends to be context dependent and at best could only be an approximation, while sentence-length tends to be under the conscious control of the writer and/or editor. For this

reason, I chose not to replicate the word and sentence length attempted in the Psychic study, which indicated no significant conclusion regarding authorship could be drawn.

The words, their frequency (total number of times they occurred) and percentage of over all distribution within the master genre corpus are as follows:

Table 5.3 Hierarchy Word Frequency List – Genre data

Genre Hierarchy Word Frequency List							
WordSmith Tools – 2/07/06 12:13:29 AM							
N	Word	Freq.	%	N	Word	Freq.	%
1	THE	17,978	6.33	2	OF	8,042	2.83
3	TO	7,473	2.63	4	A	7,450	2.62
5	AND	6,964	2.45	6	IN	5,038	1.77
7	IS	4,016	1.41	8	THAT	3,581	1.26
9	HE	3,283	1.16	10	IT	2,750	0.97
11	NOT	2,540	0.89	12	FOR	2,417	0.85
13	WAS	2,328	0.82	14	ON	1,915	0.67
15	HIS	1,848	0.65	16	AS	1,840	0.65
17	BUT	1,791	0.63	18	WITH	1,693	0.60
19	BE	1,674	0.59	20	ARE	1,385	0.49
21	BY	1,367	0.48	22	THIS	1,345	0.47
23	HAS	1,333	0.47	24	HAVE	1,284	0.45
25	I	1,271	0.45	26	THEY	1,227	0.43
27	WHO	1,111	0.39	28	HAD	1,102	0.39
29	AT	1,070	0.38	30	AN	1,063	0.37
31	WOULD	1,004	0.35	32	WILL	990	0.35
33	THERE	942	0.33	34	FROM	935	0.33
35	SAID	924	0.33	36	PRESIDENT	910	0.32
37	ABOUT	895	0.31	38	MORE	895	0.31
39	ONE	838	0.29	40	WE	824	0.29
41	OR	803	0.28	42	WERE	794	0.28
43	DO	755	0.27	44	BEEN	738	0.26
45	YOU	731	0.26	46	ALL	714	0.25
47	WHAT	710	0.25	48	IF	686	0.24
49	THEIR	648	0.23	50	SO	612	0.22

These 50 words account for 40.3% of all the words in the master genre corpus, which falls into the realm of Burrows (1992) guidelines.

In addition, as with the Psychic study, the ‘to’, which accounts for 2.53% of the master corpus has not been tagged in this analysis to determine whether authors used it as a preposition

or an infinitive. I am unable to determine how authors apply the word within their respective texts and how that may be reflective of a characteristic marker.

Analysis Matrix – Genre data

As with the earlier psychic study, having completed the hierarchal list, I built the analysis matrix. Each individual corpus (Anonymous, Klein, Kelly-Newsweek, and Kelly-New York Times) was segmented into 5000 word blocks of text. Once segmented and saved into individual text files, WordSmith tools were used to read each file and determine the number of occurrences for each word in the hierarchy. The raw data was then entered into my matrix. Below is an abbreviated sample of my matrix, the first four words in the hierarchy as well as the segment length when I converted the 5000 word block into a text file. The complete Matrix can be found in Appendix B.

Table 5.4 Sample Matrix Raw Data – Genre data

Sample Matrix Raw Data – Genre Data

Rank			1	2	3	4	Segment
Author	Texts	Segment	the	of	to	a	Length
Anonymous	Primary Colors	1	204	90	90	150	5042
Anonymous	Primary Colors	2	227	94	91	138	4993
Anonymous	Primary Colors	3	195	78	102	122	5070
Anonymous	Primary Colors	4	224	56	114	137	5015
Anonymous	Primary Colors	5	235	84	116	114	5042
Anonymous	Primary Colors	6	250	61	116	111	5089
Anonymous	Primary Colors	7	206	73	109	108	5025
Anonymous	Primary Colors	8	215	67	130	81	5063
Anonymous	Primary Colors	9	242	96	104	113	5068
Anonymous	Primary Colors	10	177	70	89	133	5015
Anonymous	Primary Colors	11	189	76	131	110	5025

The segment length is used because no complete text will divide evenly into 5000 word segments and therefore prior to analysis we need to convert the raw frequency of the words in each segment to relative frequencies for accurate statistical analysis.

Once the raw data is compiled for the analysis, the entire matrix is converted to reflect relative frequencies. As noted earlier, this matrix is used for multivariate analysis. A sample matrix of the relative frequency follows:

Table 5.5 Sample Matrix Relative Frequencies – Genre data

Sample Matrix Relative Frequencies – Genre Data

Author	Segment	the	of	to	a	and	Segment Length
Anonymous	1	0.0404601	0.01785006	0.01785006	0.029750099	0.0251884	5042
Anonymous	1	0.0454636	0.018826357	0.018225516	0.027638694	0.0166233	4993
Anonymous	1	0.0384615	0.015384615	0.020118343	0.024063116	0.0201183	5070
Anonymous	1	0.044666	0.0111665	0.022731805	0.027318046	0.0161515	5015
Anonymous	1	0.0466085	0.016660056	0.023006743	0.022610075	0.0236017	5042
Anonymous	1	0.0491256	0.011986638	0.022794262	0.021811751	0.0224013	5089
Anonymous	1	0.040995	0.014527363	0.021691542	0.021492537	0.0232836	5025
Anonymous	1	0.0424649	0.013233261	0.025676476	0.01599842	0.0175785	5063
Anonymous	1	0.0477506	0.018942384	0.020520916	0.022296764	0.0215075	5068
Anonymous	1	0.0352941	0.013958126	0.01774676	0.026520439	0.0181456	5015
Anonymous	1	0.0376119	0.015124378	0.026069652	0.021890547	0.0262687	5025

Analysis – Genre data

Again, each multivariate technique that I used (principal component analysis, discriminant analysis and cluster analysis) supplies a piece of the puzzle that ultimately completes the picture of the analysis to be explained. It is the compilation of the three techniques that allow me to draw a reasonable conclusion.

PRINCIPAL COMPONENT ANALYSIS (PCA) – To review, the goal of PCA is to reduce a large number of variables to a more manageable number of components (factors) that

allow me to make inferences about my data, as well as use them in additional statistical techniques. PCA is applied to investigate how the functions words behave among the authors and which function words become characteristic markers for individual authors.

The PCA on the Genre language data indicated that the first two components (factors) of the total variance could explain 45.8% of the variance within the data. Again, for a PCA analysis the first two components are sufficient to observe and determine which variables correlate more highly with one component or another. The first principal component accounted for 31.5% of the variance and the second factor 14.2%. Listed below is the Total Variance Explained chart from SPSS detailing component 1 and 2.

Table 5.6 Total Variance Explained – Principal Components 1 and 2 – Genre Data

Total Variance Explained – Principal Components 1 and 2 Genre Data

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	15.757	31.513	31.513	15.757	31.513	31.513
2	7.138	14.277	45.790	7.138	14.277	45.790

Extraction Method: Principal Component Analysis.

The advantage of PCA is that no mathematical assumptions are needed as the variables are allowed to choose themselves. I am allowing the fiction language of Anonymous in *Primary Colors* to find its place within the genre of the political language based on the frequency of the function words in the texts.

The first two principle components comprise the x-axis and y-axis in figure 5.1a. Figure 5.1a allows us to see how the function words behave. Variables that appear together tend to behave alike. In addition, they tend to be found more often in one group of texts and conversely less often in other group of texts. Variables found at either end of the x-axis of the plot tend to indicate authorship style. As we can see the words found on the far right of the figure 5.1a, ‘I’,

‘you’ ‘said’ and the words found to the far left of figure 5.1a, ‘the’, ‘by’, ‘president’, ‘of’, correlate most highly with component 1. Conversely, ‘be’, ‘have’, ‘but’, ‘if’ ‘is’ are found on the top of the y-axis and correlate most highly with component 2. Although there are many words toward the bottom of the y-axis on the plot, Table 5.7 indicates they did not significantly influence the variance in component 2. Initial indications in component 1, however, suggest that genre might be beginning to emerge as we begin to see hints of first person pronoun usage (signifying dialogue) that is generally found in fiction.

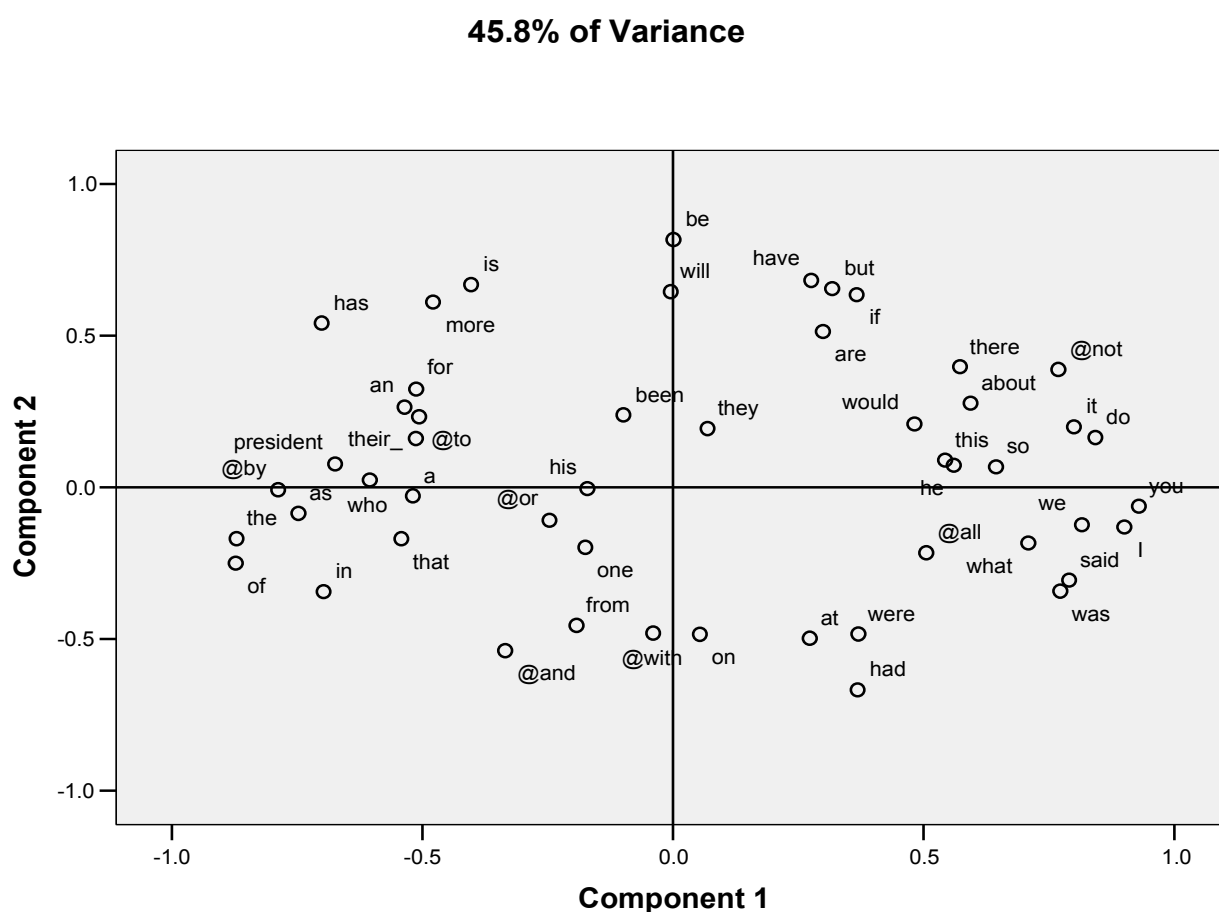


Figure 5.1a Genre Data Word Plot

Having plotted how the 50 most frequent function words behave, we need to ascertain how much of each component (1 and 2) is found in the individual segments (86 – 5000 word text segments) that represent the two author's political texts and *Primary Colors*. Figure 5.1b details where the individual segments fall within the word plot and provides a statistical group centroid that calculates where the compilation of an individual author's segments fall in the plot.

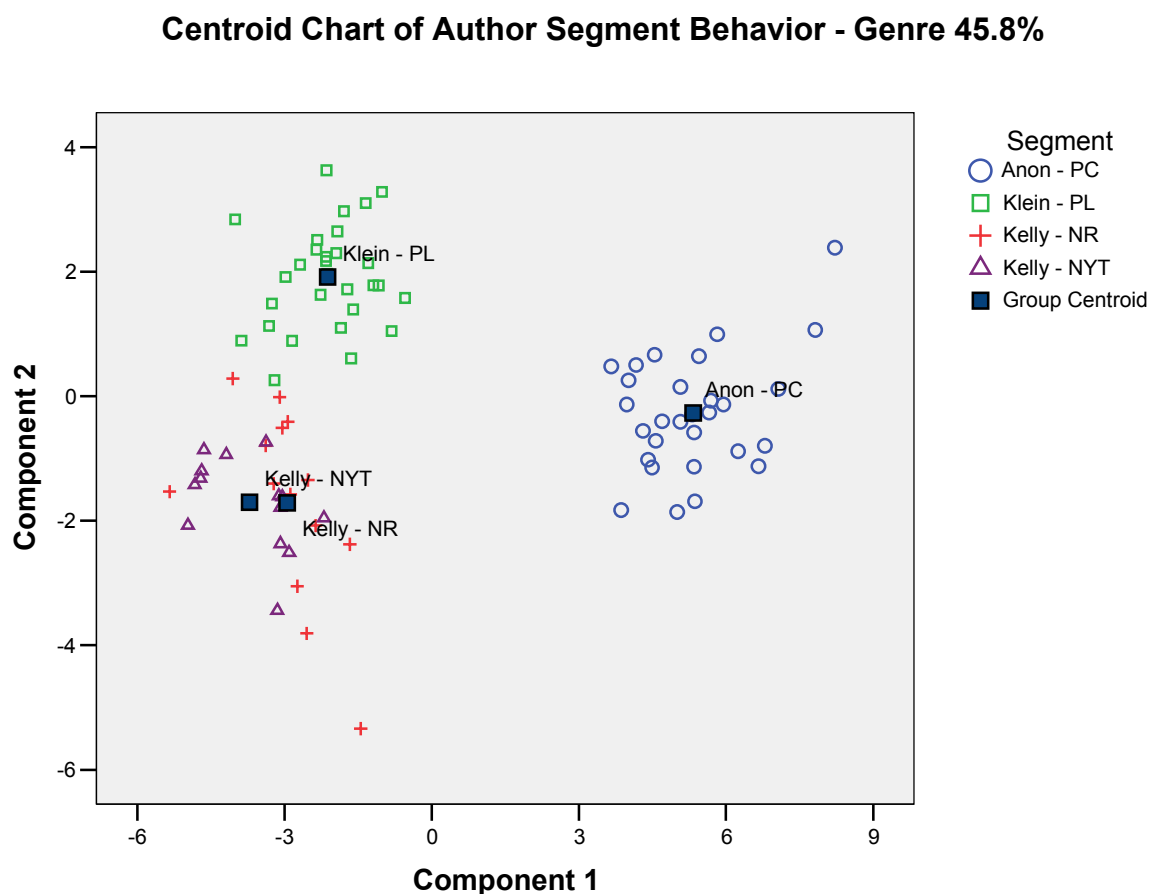


Figure 5.1b Genre Author Segment Behavior

Figure 5.1b plots the text entries by author segment and compiles them into a group centroid on the same two components found in figure 5.1a. In other words, figure 5.1a

represents a statistical picture of how the 50 function words behave, while figure 5.1b represents a statistical picture of how the actual author's texts behave based on 45.8% of the data variance.

The horizontal axis in Figure 5.1b confirms that the Genre Language study involves three distinct authors and that they clearly use the 50 function words differently. Although the function word plot (figure 5.1a) hinted that genre was beginning to emerge, the author centroid plot (figure 5.1b) clearly indicates genre plays a considerable role in how the function words are used. The political non-fiction texts of Klein and Kelly are found to the far left of the plot, while the Anonymous' fiction work resides to the far right.

It is apparent by the vertical axis in figure 5.1b that Klein and Kelly's non-fiction writings share function word usage similarities but with notable stylistic differences. Also interesting to note is that although Kelly's writings as a columnist and as a reporter share distinct similarities, even these two reflect some stylistic differences that might well be attributed to the different writing styles of those two genres.

Again, although no concrete conclusions can be drawn as to authorship of *Primary Colors* at this juncture, PCA does provide an avenue to investigate the underlying way the 50 function words behaved. The procedure provides insight as to what specific words were involved in creating the variance within the authors in the two components and the process suggests investigation to determine if there are any identifiable patterns that could explain what characteristic markers in this study best explain idiolect. Table 5.6 is a matrix of the words that correlated most highly with component 1 and component 2.

Table 5.7 Rotated Component Matrix – Genre Data**Rotated Component Matrix – Genre Data**

	Component	
	1	2
you	.932	
I	.906	
the	-.861	
of	-.858	
do	.833	
we	.821	
said	.805	
was	.790	
it	.788	
@by	-.786	
@not	.747	
has	-.729	
what	.718	
in	-.678	
president	-.677	
so	.640	
who	-.605	
be		.816
have		.696
but		.671
if		.653
is		.646
will		.644

Extraction Method: Principal Component Analysis.

Both negative and positive correlations have equal investigative value, as it is the absolute value that is considered. The closer the number is to 1, the more it contributed to the variance of the component. In the above table, the word ‘you’ correlated the highest with component 1, while the word ‘be’ correlated highest to component 2.

Unlike the Psychic language study, an analysis of the speech class each word belongs to might allow us to assign meaning as a whole. A salient pattern arises that mirrors Biber’s 1988 textual dimension and relations findings in *Variations across Speech and Writings* regarding genre usage. Biber found in one-dimension plot of genres, fiction co-occurs with many pronouns

and past tense verbs, while non-fiction co-occurs with few pronouns and few past tense verbs (p. 18). We know from the function word plot (figure 5.1a), that pronouns correlate highly with component 1 and are found on the far right of the plot. They comprise 28% of the variance found in component 1. Past tense verbs comprise 17% of the variance in component 1. Their combined representative total of 47% for component 1, fits the Biber model, indicating that we might assign the fiction genre to component 1. In turn, 67% of the variance found in component 2 can be accounted for present tense verbs and an assignment of non-fiction is a reasonable assumption. Genre obviously plays a role in determining what words accounted for the variance in these two components. The fiction/non-fiction dimensions are clearly delineated by the genre each author used.

Listed below is the table detailing the parts of speech of the words found in component 1 and component 2:

Table 5.8 Principal Components 1 and 2 Parts of Speech – Genre Data

Principal Component 1 – Parts of Speech Genre Data

Noun	Adverb	Determiner	Pronoun	Article	Preposition	Verb	Negation
president	so	what	you	the	by	do	not
		who	I		of	said	
			we		in	was	
			it			has	

Principal Component 2 – Parts of Speech Genre Data

Verbs	Conjunctions
be	but
have	if
is	
will	

A discriminant analysis and cluster analysis will be needed to determine if their function word usage varies in each genre so dramatically that the identity of Anonymous will remain a known mystery.

DISCRIMINANT ANALYSIS (DA) - The goal in applying DA is to classify the segmented texts of Klein and Kelly into uniquely defined populations and predict authorship of Anonymous. Again, to apply DA I will use the components (factors) that were developed in PCA, which reduces the large number of variables (function words) to a more manageable number of components that contain the variance within the data. Because PCA calculates as many principal components as there are variables, it is necessary to decide how many components (factors) are required to adequately represent the data in DA. Remember the goal is to explain as much of the variance as possible using as few components as possible. As with the Psychic study, I used the common criterion called the eigenvalue-greater-one, which suggests that only components with a variance of one should be included as components with a total variance of less than 1 are no better than using the individual variables. The total column under Initial Eigenvalues in Table 5.7 suggests that the first ten components are adequate for this analysis.

The ten components account for 74.3% of the total variance in the data. The remaining 40 components (25.7%) are so small that they do not significantly contribute to explaining any of the variance. Listed below is the total variance explained chart from SPSS that details the ten new components (variables) that will be used in the DA.

Table 5.9 Total Variance Explained – Principal Components 1-10, Genre Data**Total Variance Explained – Principal Components 1-10, Genre Data**

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	15.757	31.513	31.513	15.757	31.513	31.513
2	7.138	14.277	45.790	7.138	14.277	45.790
3	3.037	6.073	51.863	3.037	6.073	51.863
4	2.379	4.758	56.622	2.379	4.758	56.622
5	1.968	3.936	60.557	1.968	3.936	60.557
6	1.723	3.446	64.003	1.723	3.446	64.003
7	1.653	3.305	67.308	1.653	3.305	67.308
8	1.260	2.521	69.829	1.260	2.521	69.829
9	1.126	2.253	72.082	1.126	2.253	72.082
10	1.093	2.187	74.269	1.093	2.187	74.269
11	.968	1.936	76.205			
12	.946	1.893	78.097			
13	.875	1.749	79.847			
14	.810	1.620	81.467			
15	.731	1.461	82.928			
16	.637	1.274	84.201			
17	.611	1.222	85.423			
18	.578	1.156	86.579			
19	.551	1.102	87.681			
20	.526	1.052	88.733			
21	.493	.987	89.720			
22	.468	.936	90.656			
23	.385	.770	91.426			
24	.379	.758	92.184			
25	.354	.709	92.893			
26	.342	.684	93.577			
27	.334	.668	94.245			
28	.278	.556	94.800			
29	.274	.548	95.349			
30	.257	.514	95.863			
31	.236	.472	96.335			
32	.212	.424	96.759			
33	.189	.377	97.136			
34	.170	.340	97.476			
35	.167	.333	97.809			
36	.152	.305	98.114			
37	.127	.254	98.368			
38	.113	.226	98.594			
39	.105	.210	98.804			
40	.092	.183	98.987			
41	.084	.167	99.154			
42	.080	.159	99.313			
43	.070	.139	99.453			
44	.063	.126	99.579			
45	.056	.111	99.690			
46	.049	.097	99.787			
47	.036	.071	99.858			
48	.029	.058	99.916			
49	.023	.045	99.961			
50	.019	.039	100.000			

Extraction Method: Principal Component Analysis.

Prior to running the analysis, we must determine whether our predictors (50 function words) indicate significant differences in the means among our three authors. In other words are our function word variables good predictors of author variance. The significance column (bolded) should reflect a p value $< .05$. As table x illustrates, with the exception of nine variables, which have been bolded, the statistics indicate a significant difference in the means among the predictors used in the analysis. The words - 'to', 'for', 'was', 'his', 'one', 'we', 'been', 'there', 'so' - do not appear to differ significantly among Klein and Kelly in their non-fiction texts. The fiction text of *Primary Colors* is not used in the analysis of Table 5.10 because the DA will predict to which author the 29 text segments from the novel *Primary Colors* will gravitate.

Table 5.10 Tests of Equality of Group Means – Genre Data

Tests of Equality of Group Means – Genre Data

	Wilks' Lambda	F	df1	df2	Sig.
the	.519	25.039	2	54	.000
of	.435	35.052	2	54	.000
@to	.976	.652	2	54	.525
a	.855	4.566	2	54	.015
@and	.519	25.061	2	54	.000
in	.688	12.240	2	54	.000
is	.669	13.359	2	54	.000
that	.706	11.228	2	54	.000
he	.876	3.838	2	54	.028
it	.592	18.605	2	54	.000
@not	.423	36.787	2	54	.000
for	.931	2.006	2	54	.144
was	.939	1.751	2	54	.183
on	.752	8.899	2	54	.000
his	.934	1.899	2	54	.160
as	.643	15.007	2	54	.000
but	.388	42.636	2	54	.000
@with	.742	9.364	2	54	.000
be	.457	32.074	2	54	.000
are	.646	14.798	2	54	.000
@by	.876	3.834	2	54	.028
this	.818	5.991	2	54	.004
has	.567	20.592	2	54	.000
have	.608	17.433	2	54	.000
I	.894	3.206	2	54	.048
they	.884	3.537	2	54	.036
who	.834	5.377	2	54	.007
had	.625	16.212	2	54	.000

at	.686	12.381	2	54	.000
an	.838	5.219	2	54	.008
would	.915	2.513	2	54	.090
will	.681	12.654	2	54	.000
there	.592	18.601	2	54	.000
from	.837	5.263	2	54	.008
said	.673	13.090	2	54	.000
president	.777	7.727	2	54	.001
about	.746	9.176	2	54	.000
more	.488	28.282	2	54	.000
one	.942	1.672	2	54	.197
we	.981	.513	2	54	.602
@or	.705	11.315	2	54	.000
were	.831	5.502	2	54	.007
do	.804	6.565	2	54	.003
been	.899	3.030	2	54	.057
you	.973	.759	2	54	.473
@all	.754	8.790	2	54	.000
what	.861	4.356	2	54	.018
if	.395	41.412	2	54	.000
their_	.904	2.875	2	54	.065
so	.900	3.002	2	54	.058

In addition to testing the viability of the predictors (variables), we want to confirm whether there are significant differences between Klein and Kelly across the function words. In other words, are the two authors using the variables differently? The Wilks' Lambda test predicts whether the predictors (variables) differentiated significantly among the authors. This test is significant at $p < .05$. Table 5.11 indicates there are significant differences among the authors across a compilation of all predictor variables in the two functions that DA uses.

Table 5.11 Wilks' Lambda – Genre data

Wilks' Lambda – Genre Data

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.000	233.783	100	.000
2	.033	100.932	49	.000

Having confirmed that the variables and that Klein and Kelly have significant differences, I proceeded to the classification results of their individual segments of texts and the 29 segments of *Primary Colors*. Table 5.12 indicates that of the original group cases (segmented texts of Klein and Kelly) Klein's "Public Lives" texts and Kelly's *New York Times* articles

classified correctly. However, the DA found that two of Kelly's text segments from his *New Republic* column shared more similarities with his *New York Times* articles and reclassified.

For the 29 *Primary Colors* text segments, SPSS has computed discriminant scores and predicted group membership based on the Klein and Kelly's non-fiction segmented texts. Listed below are the classification results from the DA:

Table 5.12 Classification Results – Genre data
Classification Results(b)

		Predicted Group Membership			
	Segment	Klein - PL	Kelly - NR	Kelly - NYT	Total
Original	Count Klein - PL	29	0	0	29
	Kelly - NR	0	12	2	14
	Kelly - NYT	0	0	14	14
	Ungrouped Cases	23	6	0	29
%	Klein - PL	100.0	.0	.0	100.0
	Kelly - NR	.0	85.7	14.3	100.0
	Kelly - NYT	.0	.0	100.0	100.0
	Ungrouped cases	79.3	20.7	.0	100.0

b. 96.5% of original grouped cases correctly classified.

My DA analysis of the language indicates that of the 29 text segments from *Primary Colors*, there is a 79.3% probability that 23 of the segments reflect the non-fiction language of Joe Klein from his political "Public Lives" columns that he wrote during the same time period as *Primary Colors*. There is a 20.7% probability that the remaining six segments in question reflect the language of Michael Kelly. It is notable, however, that the six segments attributed to Michael Kelly are most similar to his political column texts at *The New Republic*, not his texts as a reporter from *The New York Times*. Apparently, the underlying variation of the fiction novel is more similar to a column environment than to a reporter's writings in an article genre. This confirms what figure 5.1b details in the PCA author plot that although Kelly's writings as a columnist and as a reporter share distinct similarities, even these two genres reflect some stylistic

differences. The differences are enough to attribute the six segments to just one genre of his writings.

Figure 5.1c visually illustrates how the Klein and Kelly political, non-fiction text segments respectively group, and how the 29 *Primary Colors* text segments behave in the environment of the analysis. Despite the genre difference between fiction and non-fiction, DA was able to attribute authorship of Anonymous to Joe Klein based on his characteristic usage of the function words. Clearly, genre influences the patterning of language; however, this analysis indicates that our subconscious use of function words is not significantly affected by the text genre we choose. Our individual markers can be teased out of the data by a quantitative statistical investigation.

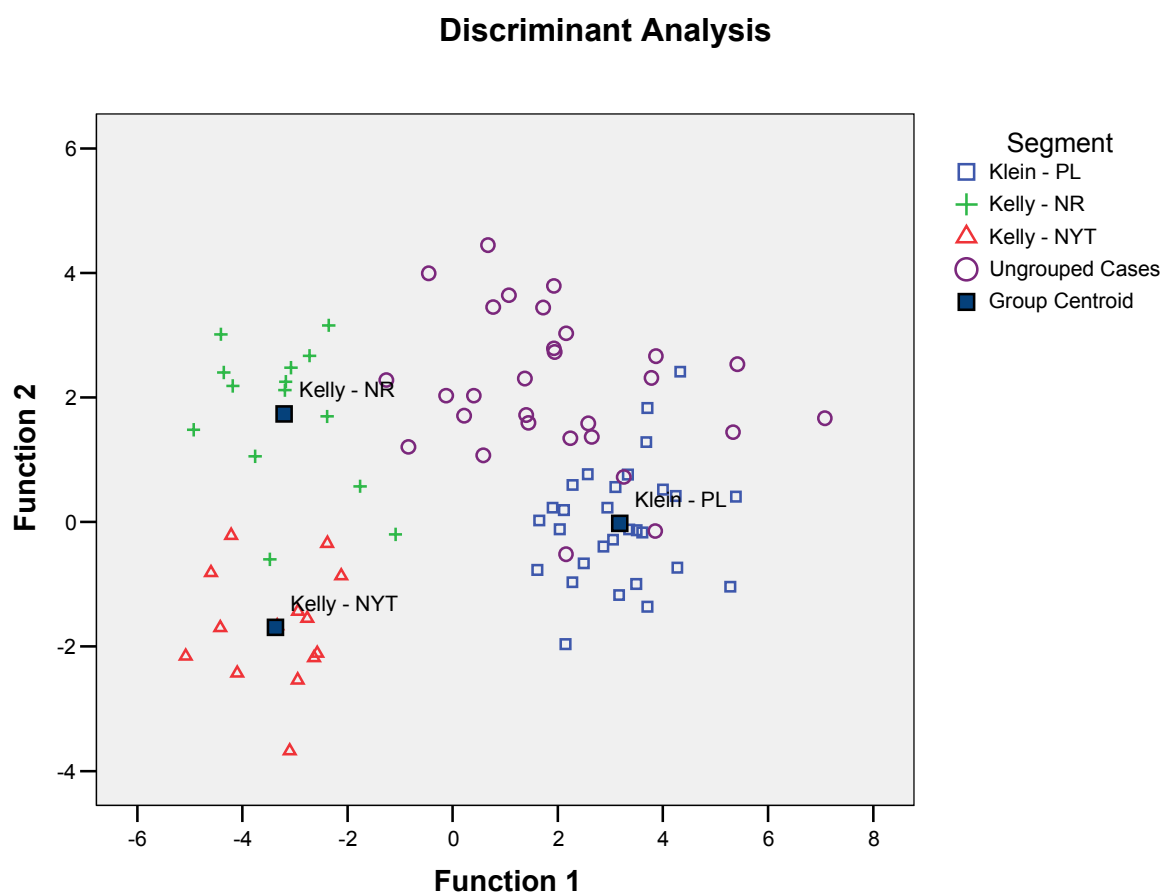


Figure 5.1c Genre Discriminant Analysis

CLUSTER ANALYSIS (CA) – To review, whereas DA classified the 29 text segments of the novel *Primary Colors* into predicted group membership, my goal in using CA are that I do not have to make any underlying assumptions about the data. Statistics will allow the groups to form based on the characteristics of the individual segments of each author including the 29 segments of the novel by Anonymous. I am operating under the assumption that I do not know from which subgroup (Klein – PL, Kelly – NR, Kelly – NYT and Anonymous) the individual segments originate.

As with the Psychic Language study, I again used Hierarchical Clustering, which is one of the most straightforward methods in CA. It begins with each case (segment) being a cluster unto itself and at each successive step, similar clusters are merged. As explained, dendrograms are read from left to right. Figure 5.1d represents the findings of my CA analysis on the 86 individual segments delineated by author genre subgroup. Immediately apparent is that the majority of each author's individual segments tend to cluster with themselves indicating the uniqueness of their individual language in the usage of function words. For example, Joe Klein's segments 45, 48, and 54 identify most closely with each other on the first vertical line on the left and then cluster with his segments 42, 43 and 50. Similar clusters continue to merge and project similarity among the individual segments until the subgroups are exhausted. It is notable that of the 29 segments belonging to Anonymous, 28 of them initially cluster with the Joe Klein segments rather than segments from Michael Kelly. On the seventh vertical line, we see that a large portion of the sub-group of Anonymous clearly clusters itself with the majority of Joe Klein's subgroup prior to attempting to align itself with Michael Kelly. Without any underlying assumptions, the characteristics of the individual segments confirm that Anonymous identifies with its own segments originally, indicating a genre issue, but when the segments are allowed to

further align they consider themselves most similar to Joe Klein despite the difference in genre.

As with Discriminant Analysis, authorship of *Primary Colors* points to Joe Klein even in a cross genre environment.

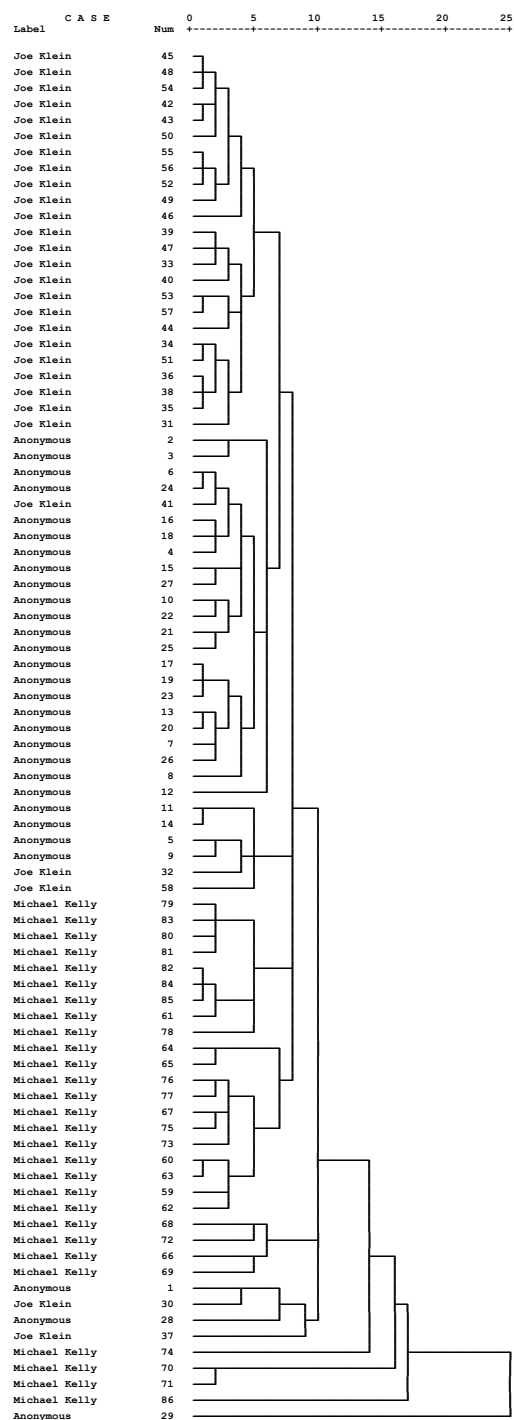


Figure 5.1d Genre Cluster Analysis

Unlike the Psychic language study, there is no real question as to authorship of the roman à clef about President Bill Clinton's 1992 presidential campaign. Joe Klein admitted to being author. This case study, however, does confirm that the quantitative multivariate triad methodology can successfully cross genre lines and predict authorship to a high probability. Joe Klein's function word usage, which contains his individual idiolect markers, is preserved when crossing over from the genre of fiction to non-fiction political texts. His individual idiolect contains habitual characteristic markers that simply cannot be easily concealed when the text segments of the novel *Primary Colors* are allowed to find their place within text segments containing his language from another genre.

Even though we know that texts in different genres (political non-fiction versus political fiction) by one author might differ more than texts by different authors in the same genre the subconscious use of function words appear to reflect deeply ingrained linguistic habit that cannot be hidden by a change in register. Genre as we saw clearly delineates itself in style, but does not serve as an exclusive barrier in situations where authorship is in question.

CHAPTER 6

CONCLUSION

In this quantitative statistical investigation, I analyze two different author scenarios that entail language concealment. I focus specifically on whether individuals can disguise their individual idiolects by changing the characteristics of their writing style, deliberately or unintentionally, or whether an individual's voice is so habitually ingrained with style markers that it can be identified even if disguised. The investigation stems from my hypothesis that an individual's idiolect contains characteristic markers, which can be used in scientific linguistic quantitative investigations to determine the authorship of various written documents.

I designed each case study to answer specific questions about authorship identification using quantitative statistical multivariate methodology. A specific mindset is necessary when using extensive statistical methodology and it bears consideration, not only from a researcher's perspective but also from the reader's perspective:

“A statistically significant result never explains the quantitative distribution under study – the analyst is always responsible for explanations – but suggests to the analyst which distributions are probably worthy of explanation” (Kretzschmar, Meyer, & Ingegneri, 1997).

In other words, the results in this type of analysis are not in the numbers themselves but in the explanation of the numbers. Each multivariate technique that I used (principal component analysis, discriminant analysis, and cluster analysis) supplies a piece of the puzzle that ultimately completes the picture of the analysis to be explained. The compilation analyses of the three techniques (multivariate triad) used in each of the studies, allows me to draw the conclusion that

the methodology used in conjunction with high frequency function words is robust enough to predict authorship to a high probability.

Principal component analysis allows us to see how the 50 function words behave overall and distinguishes for us how the actual author's texts behave using those 50 function words. It confirms the distinctiveness of the authors in the study. In addition, principal component analysis provides insight as to what specific words were involved in creating the variance within authors and suggests further investigation to determine if there are any identifiable patterns that can explain what characteristic markers best explain idiolect. Discriminant analysis classifies the authors into uniquely defined populations and predicts authorship of the investigated text. Prior to running the DA, it allows us to check whether the predictors indicate significant differences in the means among the authors – determining if the function word variables are good predictors of author variance. In addition, DA confirms whether there are significant differences among the authors across the function words. Finally, cluster analysis classifies all the texts into groups based on the characteristics of the individual segments of each author and allows us to determine the predicted group authorship of the investigated text. I cannot emphasize enough, that it is the compilation of these techniques that allows me to discuss to what degree the findings of each individual study are able to answer the questions postulated in the introduction of this dissertation. I refer to them as the multivariate analysis triad.

Chapter 4 entailed a renowned psychic, Ruth Montgomery (1971) who claimed that her deceased friend, sensitive Arthur Ford, was continuing to speak through her in the form of automatic text writings and dictated an entire book, *A World Beyond*. Ruth Montgomery said that she firmly believed that the voice in *A World Beyond* was not her own. I questioned if a

quantitative linguistic analysis could determine whose voice found in *A World Beyond* and substantiate psychic channeling claims.

I consider this an unintentional case of language concealment. Ruth Montgomery was not a fly-by-night, street corner stall psychic, attempting to eek out a meager crystal-ball living from those that needed unsubstantiated comfort. She was considered a top White House correspondent, covering every administration from President Roosevelt through President Johnson prior to discovering her self-proclaimed talent of channeling spirits from the other side - which led her to write 15 best-selling books on paranormal topics. Her subsequent faith in her colleague, noted sensitive Arthur Ford was resolute (Montgomery, 1999). Although many skeptics would disagree, I believe she truly had faith in her psychic abilities and that her claim, “This book I believe to be Arthur Ford’s own account of the life in the next stages of existence beyond the portal man calls death” was a sincere one (Montgomery, 1971, p. 3).

Unfortunately, despite her sincerity, the statistical multivariate analysis does not bear out her claim. The multivariate analysis triad indicates that there is minimal probability that Arthur Ford’s spirit wrote the unknown text and it suggests that he is not a potential match for the nine unknown segments of text under investigation. Further, there is an 88.9% probability that the language in *The World Beyond* belongs to Ruth Montgomery. Ruth Montgomery’s individual idiolect contains habitual characteristic markers that simply cannot be concealed when the unknown text is allowed to find its place within the language based on the frequency of function word usage.

Chapter 5 entailed a situation whereby Joe Klein, a well know political columnist and television political commentator, wrote the political novel *Primary Colors* under the pseudonym Anonymous to disguise his identity in a highly charged political environment. His deliberate

attempt to conceal his identity as the writer of this unflattering *roman à clef* about President Bill Clinton's 1992 presidential campaign, consumed the world of United States politics for seven months. I questioned what happens to the individual's voice in a pseudonymic environment and does it retain its habitual markers. I also questioned if the fact that Joe Klein's writings cross genres, from political essay to political novel, does an analysis make it more difficult to ascertain his authorship.

The multivariate analysis triad clearly indicates that in a pseudonymic environment initially texts in different genres (political writings versus political fiction) by one author might differ more than texts by different authors in the same genre. Plot 4.1b indicates that Klein's usage of function words differ dramatically across genre. However, further analysis reveals that a quantitative statistical study can successfully cross genre lines and predict authorship to a high probability as illustrated in plot 4.1c. The analyses reveals that of the 29 text segments from *Primary Colors*, there is a 79.3% probability that 23 of the segments reflect the non-fiction language of Joe Klein from his political "Public Lives" columns that he wrote during the same time period as *Primary Colors*. Joe Klein's function word usage, which contains his individual idiolect markers, is preserved when crossing over from the genre of political fiction to political non-fiction texts. His individual idiolect contains habitual characteristic markers that cannot be easily concealed when the text segments of the novel *Primary Colors* are allowed to find their place among other texts containing his language from another genre despite his deliberate attempt to disguise his voice.

My findings indicate that both of these studies successfully identify to a high degree of probability authorship and that they can be replicated if significant amounts of text are available.

Clearly, this multivariate triad methodology has value in a forensic arena where considerable amounts of text exist upon which a linguist can apply the triad multivariate technique.

In a forensic environment, however, rarely do problems occur that would supply an analyst with substantial amounts of text. The JonBenét Ramsey note was comprised of 371 words. Both Don Foster and Gerald McMennamin were given unprecedented access to the suspect parties and/or additional suspect writings upon which they could perform a comparative qualitative stylistic analysis. Unfortunately, I believe that the variant approaches involved in the analysis methodology could not provide a verifiable result. Stylistic analyses do have their place in Forensic linguistics. They can provide an investigation with critical data about a potential suspect. I maintain, however, that rarely does the data from a purely qualitative analysis qualify as evidence in an author identification scenario but that it should be considered an extremely valuable profiling tool upon which further established quantitative methodology should be applied.

Even with the unprecedented access given to Foster and McMennamin, however, any multivariate analysis is predicated on having as many cases (text segments) as it has variables (function words). That suggests that it is highly unlikely that the statistical triad multivariate technique in my study would be able to identify potential authorship, given the parameters that I used in the Psychic language and Genre study in this dissertation. Sample size becomes the issue - Baayen in *Word Frequency Distribution* cautions us that sample size is critical in any statistical analysis of language data, as the results can vary dramatically (2001).

Despite the suggestion that the multivariate analysis triad would be ineffective and present limitations in a forensic environment, I am confident of the robustness of multivariate analysis using high frequency function words. I propose that additional studies that involve

drilling down into these original two studies, using less and less text as well as fewer function words, will help to determine at what level multivariate analyses will begin to lose its predictive probability value. At the level prior to its failure, I recommend that supplementary studies be performed on retrospective, resolved cases such as the Unabomber case, to determine the efficacy of the methodology.

These types of studies, if replicable and significant, would provide a quantitative procedure upon which linguists can approach a forensic problem. Used in conjunction with a stylistic analysis, the quantitative approach would provide a consistent method that analyzes potential rates of error, and presents our court systems with an acceptable technique that qualifies under rule 702 of the Federal Rules of admissible evidence cited in *Daubert v. Merrell Dow Pharmaceuticals* (1993).

This dissertation establishes that authorship can be determined to a high probability by employing a triad of multivariate analyses that utilizes function words. I concur, however, with Burrows (2003) who cautions that we have not yet “reached “the holy grail of stylometry”, but that multivariate methods “are increasingly reliable, our use of them is ever more rigorous, and we have vast new corpora to strengthen our comparisons” (p. 7). Both of my studies illustrate that language can be quantitatively analyzed to determine authorship however, as promising as the results are, it is just the beginning and quantitative author identification requires further replicable research especially in cases that involve minimal sample size data.

I further contend the evidence in these two studies does not transcend the myth that this analysis constitutes “literary DNA” or that it lends itself to conclude quantitatively that an individual’s language possesses an identifiable “fingerprint” as has been suggested. Multivariate statistical techniques provide us with probabilities not absolutes.

I respectfully provide a realistic expectation as to what a linguist can quantitatively offer under the rubric of forensic linguistics - author identification in situations that entail significant amounts of text and/or with authors that navigate across genre lines. I candidly invite future studies of various sample size where linguists can with consensus and credibility, discern potential authorship in this ubiquitous world of “forensics” whereby the evidentiary data we furnish as linguists begins to meet the standards of evidence outlined by our court system.

BIBLIOGRAPHY

- AIM. (2000). Optical Character Recognition, *The Association for Automatic Identification and Data Capture Technologies* (pp. 1 - 10). Pittsburgh, PA: AIM, Inc.
- Allen, M. (2004). *Should you use a Pseudonym?*, from www.writing-world.com
- Anonymous. (1996). *Primary Colors*. New York: Random House.
- AtlanticMedia. (2004). *Michael Kelly Biography*, from http://kellyaward.com/mk_about_mk.html
- Baayen, H. (2001). *Word Frequency Distribution*. Boston: Kluwer Academic.
- Baayen, H., Van Halteren, H., & Tweedie, F. (1996). Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution. *Literary and Linguistic Computing*, 11(3), 121-131.
- Bailey, R. W. (1979). Author Attribution in a Forensic Setting. In D. E. Ager, F. E. Knowles & J. Smith (Eds.), *Advances in Computer-aided Literary and Linguistic Research*. Birmingham: AMLC.
- Biber, D. (1988). *Variation across speech and writing*. New York: Cambridge University Press.
- Biber, D. (1989). A Typology of English Texts. *Linguistics*, 27, 3-43.
- Biber, D. (1994). An Analytical Framework for Register Studies. In D. Biber & E. Finegan (Eds.), *Sociolinguistic Perspectives on Register* (pp. 31-56). New York: Oxford University Press.
- Biber, D. (1995). *Dimensions of Register Variation*. Cambridge: Cambridge University Press.
- Biber, D., & Finegan, E. (1989). Drift and the Evolution of English Style: A History of Three Genres. *Language*, 65(65), 487-517.

- Binongo, J. N. G. (1994). Joaquin's Joaquinquerie, Joaquinquerie's Joaquin: A statistical Expression of a Filipino Writer's Style. *Literary and Linguistic Computing*, 9(4), 267-279.
- Browne, S. (2002). *Conversations from the Other Side*. Carlsbad, California: Hay House, Inc.
- Burrows, J. F. (1987). Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style. *Literary and Linguistic Computing*, 2(2), 61-70.
- Burrows, J. F. (1989). 'An Ocean Where Each Kind ... ': Statistical Analysis and Some Major Determinants of Literary Style. *Computers and the Humanities*, 23(4-5), 309-321.
- Burrows, J. F. (1992a). Computers and the Study of Literature. In C. S. Butler (Ed.), *Computers and Written Texts*. Oxford: Blackwell.
- Burrows, J. F. (1992b). Not Unless You Ask Nicely: The Interpretative Nexus between Analysis and Information. *Literary and Linguistic Computing*, 7(2), 91-109.
- Burrows, J. F. (2003). Questions of Authorship: Attribution and Beyond. *Computers and the Humanities*, 37, 5-32.
- Canada, M. (2006). *All American Glossary of Literary Terms*. Retrieved January 9, 2006, from <http://www.uncp.edu/home/canada/work/allam/general/glossary/htm>
- Carvajal, M. (July 18, 1996). Columnist's Mea Culpa: I'm Anonymous. *The New York Times*, p. 23.
- Chaski, C. E. (2001). Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, 8(1), 1-65.
- 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(1), 103-113.

- Craig, H. (2000). Is the Author Really Dead? An Empirical Study of Authorship in English Renaissance Drama. *Empirical Studies of the Arts*, 18(2), 119-134.
- Craig, H., Burrows, John. (2001). Lucy Hutchinson and the authorship of two seventeenth century poems: a computational approach. *The Seventeenth Century*, 16, 259-282.
- Ferguson, C. A. (1994). Dialect, Register, and Genre: Working Assumptions about Conventionalization. In D. Biber & E. Finegan (Eds.), *Sociolinguistic Perspectives on Register*. New York: Oxford University Press.
- Ford, A. (1953). *Why We Survive*. Cooksburg, N.Y.: Gutenberg Press.
- Ford, A. (1968). *Unknown but Known*. New York: Signet Books.
- Foster, D. (2000). *Author Unknown: On the Trail of Anonymous*. New York: Henry Holt and Company.
- Glass, S. (1998, Feb). Prophets and Losses. *Harper's*, 296, 69-72.
- Grant, T., & Baker, K. (2001). Identifying reliable , valid markers of authorship: a response to Chaski. *Forensic Linguistics*, 8(1), 66-79.
- Green, S. S., Neil J. (2003). *Using SPSS for Window and Macintosh*. Upper Saddle River: Prentice Hall.
- Hockey, S. (2000). *Electronic Texts in the Humanities*. New York: Oxford University Press.
- Holmes, D. (1994). Review: Attributing Authorship: An Introduction. *Computers and the Humanities*, 28, 87-106.
- Holmes, D. I. (1994). Authorship Attribution. *Computers and the Humanities*, 28(2), 87-106.
- Holmes, D. I. (1998). The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing*, 13(3), 111-117.

- Holmes, D. I., & Forsyth, R. S. (1995). The Federalist Revisited: New Directions in Authorship Attribution. *Literary and Linguistic Computing*, 10(2), 111-127.
- Hoover, D. I. (2001). Statistical Stylistics and Authorship Attribution: An Empirical Investigation. *Literary and Linguistic Computing*, 16(4), 421-444.
- James, C. (1996, July 21, 1996). Anonymous Shows His Colors. *The New York Times*, p. 2.
- Jaynes, J. (1976). *The origin of consciousness in the breakdown of the bicameral mind*. Boston: Houghton Mifflin.
- JOHNSESL. (2006). *Common Contractions*. Retrieved February 8, 2006, from <http://www.johnsesl.com>
- Johnson, D. E. (1998). *Applied Multivariate Methods for Data Analysts*. Pacific Grove: Duxbury Press.
- Kelly, M. (1991a). Amman Diarist. *New Republic*, 204(8), 42.
- Kelly, M. (1991b). Back to the Hills. *New Republic*, 204(22), 23-26.
- Kelly, M. (1991c). Before the Storm. *New Republic*, 204(5), 21-23.
- Kelly, M. (1991d). Blitzed. *New Republic*, 204(6), 21-22.
- Kelly, M. (1991e). Desert Rat. *New Republic*, 204(7), 14-18.
- Kelly, M. (1991f). Highway to Hell. *New Republic*, 204(13), 11-14.
- Kelly, M. (1991g). Kiss of Victory. *New Republic*, 204(11), 18-21.
- Kelly, M. (1991h). The other Hell. *New Republic*, 204(19), 14-16.
- Kelly, M. (1991i). The Rape and Rescue of Kuwait City. *New Republic*, 204(12), 20-25.
- Kelly, M. (1991j). Rolls-Royce Revolutionaries. *New Republic*, 204(14), 22-24.
- Kelly, M. (1991k). Speech Defect. *New Republic*, 204(9), 16-17.

- Kelly, M. (1992a). Altering course, Clinton renews draft defensive. (Cover story). *New York Times*, 141(49090), A1.
- Kelly, M. (1992b). As race looks tighter, theme is truth and trust. (Cover story). *New York Times*, 142(49134), A1.
- Kelly, M. (1992c). Being whatever it takes to win election. *New York Times*, 141(49067), 1.
- Kelly, M. (1992d). Bush holds back in digs at Clinton over draft issue. (Cover story). *New York Times*, 141(49091), A1.
- Kelly, M. (1992e). The candidates as culture vultures. *New York Times*, 141(49025), 1.
- Kelly, M. (1992f). The center of attention. (Cover story). *New York Times*, 142(49107), A1.
- Kelly, M. (1992g). Clinton and Bush compete to be champion of change. (Cover story). *New York Times*, 142(49136), 1.
- Kelly, M. (1992h). Clinton and Bush in a sprint as race for White House ends. (Cover story). *New York Times*, 142(49139), A1.
- Kelly, M. (1992i). Clinton foresees Bush trimming Social Security. (Cover story). *New York Times*, 141(49079), A1.
- Kelly, M. (1992j). Clinton may carry the campaign into office. (Cover story). *New York Times*, 142(49166), A1.
- Kelly, M. (1992k). Clinton says Bush imperils elderly. (Cover story). *New York Times*, 141(49077), A1.
- Kelly, M. (1992l). Clinton, after raising hopes, tries to lower expectations. (Cover story). *New York Times*, 142(49145), A1.
- Kelly, M. (1992m). Clinton, calling economy weak, intends to focus on its problems. (Cover story). *New York Times*, 142(49146), A1.

- Kelly, M. (1992n). Clinton, sketching plan for economy, counsels patience. (Cover story). *New York Times*, 142(49149), A1.
- Kelly, M. (1992o). Clinton's camp says it is wary of Perot inroads. (Cover story). *New York Times*, 142(49127), A1.
- Kelly, M. (1992p). Contest of two generations has risks for both nominees. (Cover story). *New York Times*, 141(49074), 1.
- Kelly, M. (1992q). Debating done, Bush and Clinton begin final push. (Cover story). *New York Times*, 142(49126), A1.
- Kelly, M. (1992r). Encircling Arkansas, Bush opens harsh attack on Clinton's record. (Cover story). *New York Times*, 142(49098), A1.
- Kelly, M. (1992s). The making of a first family: A blueprint. (Cover story). *New York Times*, 142(49150), 1.
- Kelly, M. (1992t). Merely mortal. *New York Times*, 142(49102), 1.
- Kelly, M. (1992u). Though advisers differ, Clinton's in tune with all. (Cover story). *New York Times*, 141(49088), 1.
- Kelly, M. (1993a). Clinton is greeted with a major test on Baird hearings. (Cover story). *New York Times*, 142(49219), A1.
- Kelly, M. (1993b). The Game. *New York Times Magazine*, 143(49501), 62.
- Kelly, M. (1993c). The Guinier affair aggravates Clinton's credibility problem. *New York Times*, 142(49354), 1.
- Kelly, M. (1993d). Hillary Clinton takes to Hill in vivid display of influence. (Cover story). *New York Times*, 142(49233), A1.

- Kelly, M. (1993e). Household hiring is trickier with new broom in Capital. (Cover story). *New York Times*, 142(49240), A1.
- Kelly, M. (1993f). The new-politics myth. (Cover story). *New York Times*, 142(49354), 1.
- Kelly, M. (1993g). President's early troubles rooted in party's old strains. (Cover story). *New York Times*, 142(49230), A1.
- Kelly, M. (1993h). Read the fine print to know if a promise really counts. (Cover story). *New York Times*, 142(49223), A1.
- Kelly, M. (1994a). Howards End. *New Republic*, 210(12), 11-12.
- Kelly, M. (1994b). The president's past. (Cover story). *New York Times Magazine*, 143(49774), 20.
- Kelly, M. (1996a). Dangerous minds. *New Republic*, 215(27), 6.
- Kelly, M. (1996b, 1996/12/23/). Friends like these. *New Republic*, p. 6.
- Kelly, M. (1996c). Our hero. *New Republic*, 215(23), 6.
- Kelly, M. (1996d). A plea for diversity. *New Republic*, 215(25), 6.
- Kelly, M. (1996e). The script. *New Republic*, 215(24), 6.
- Kelly, M. (1997a). Banality and evil. *New Republic*, 216(18), 6.
- Kelly, M. (1997b). Breach of promise. *New Republic*, 216(17), 4-49.
- Kelly, M. (1997c). The China syndrome. *New Republic*, 217(4), 4.
- Kelly, M. (1997d). CITIC-VIP. *New Republic*, 216(10), 4-49.
- Kelly, M. (1997e). Class. *New Republic*, 216(24), 6-45.
- Kelly, M. (1997f). Curiouser. *New Republic*, 216(7), 6.
- Kelly, M. (1997g). Feint victory. *New Republic*, 216(22), 6.
- Kelly, M. (1997h, 1997/03/03/). Follow the money. *New Republic*, p. 4.

- Kelly, M. (1997i). For America's sake. *New Republic*, 216(12), 4.
- Kelly, M. (1997j). The freelance. *New Republic*, 217(8), 6.
- Kelly, M. (1997k). Geronimo! *New Republic*, 217(6/7), 6.
- Kelly, M. (1997l). The goodwill scam. *New Republic*, 216(1/2), 6.
- Kelly, M. (1997m). The great divider. *New Republic*, 217(1), 6-41.
- Kelly, M. (1997n, 1997/05/19/). I volunteer. *New Republic*, p. 4.
- Kelly, M. (1997o). In principle. *New Republic*, 216(23), 6-49.
- Kelly, M. (1997p). In the pipeline. *New Republic*, 216(25), 6-45.
- Kelly, M. (1997q). Judge dread. *New Republic*, 216(13), 6.
- Kelly, M. (1997r). Op city. *New Republic*, 217(10/11), 6-45.
- Kelly, M. (1997s). Plan B. *New Republic*, 216(14), 4.
- Kelly, M. (1997t). Presidential. *New Republic*, 216(8), 6.
- Kelly, M. (1997u). A promise kept. *New Republic*, 217(9), 6-41.
- Kelly, M. (1997v). The Reich stuff. *New Republic*, 216(26), 6-42.
- Kelly, M. (1997w, 1997/03/17/). Right and righteous. *New Republic*, p. 4.
- Kelly, M. (1997x). Rope-a-hope. (Cover story). *New Republic*, 216(6), 6.
- Kelly, M. (1997y). Take your medicine. *New Republic*, 217(2/3), 6-45.
- Kelly, M. (1997z). Taking sides. *New Republic*, 217(12), 4.
- Kelly, M. (1997{). A toast. *New Republic*, 216(15), 6-49.
- Kelly, M. (1997|). Total war. *New Republic*, 216(5), 6-49.
- Kelly, M. (1997}, 1997/04/21/). Webb site. *New Republic*, p. 4.
- Kelly, M. (1997~). What news? *New Republic*, 217(5), 6.
- Kelly, M. (1997□, 1997/05/12/). What, me worry? *New Republic*, pp. 6-49.

- Kelly, M. (1997€). Why it matters (II). *New Republic*, 216(4), 6.
- Kelly, M. (1997□). Why it matters. *New Republic*, 216(3), 6-41.
- Kelly, M. (1997,). Wonk new world. *New Republic*, 216(21), 6.
- Kelly, M., & Dillon, S. (1993). Two Hillary Clintons: One trailblazing, one traditional. (Cover story). *New York Times*, 142(49224), A1.
- Kelly, M., & Johnston, D. (1992). Campaign renews disputes of the Vietnam War years. (Cover story). *New York Times*, 142(49114), A1.
- Kennedy, H., & Kornblut, A. E. (February 16, 1996). Newsweek Scribe 'Primary' Suspect. *The New York Daily News*, p. 2.
- King, S. (2004). from www.stephenking.com
- Klein, J. (1992a). A blue Christmas for Elvis. *Newsweek*, 120(26), 33.
- Klein, J. (1992b). Bush: Praying for rain. *Newsweek*, 120(13), 23.
- Klein, J. (1992c). Bush's desperate game. (Cover story). *Newsweek*, 120(16), 26.
- Klein, J. (1992d). Chilly scenes of winter. *Newsweek*, 120(25), 42.
- Klein, J. (1992e). Clinton and the Tao of BAU. *Newsweek*, 120(22), 38.
- Klein, J. (1992f). Conundrum in the classroom. *Newsweek*, 120(11), 32.
- Klein, J. (1992g). Copping a domestic agenda. *Newsweek*, 120(23), 29.
- Klein, J. (1992h). Fighting the squish factor. *Newsweek*, 120(10), 39.
- Klein, J. (1992i). Little lies and big whoppers. *Newsweek*, 120(9), 36.
- Klein, J. (1992j). Perot's people: Second thoughts. *Newsweek*, 120(14), 44.
- Klein, J. (1992k). Prisoner of the people. *Newsweek*, 120(18), 58.
- Klein, J. (1992l). The relentless suitor. *Newsweek*, 120(4), 34.
- Klein, J. (1992m). Walking small. *Newsweek*, 120(5), 29.

- Klein, J. (1992n). When everyone's an amateur. *Newsweek*, 120(24), 42.
- Klein, J. (1992o). The year of living seriously. *Newsweek*, 120(15), 45.
- Klein, J. (1993a). Bungee-jumping. *Newsweek*, 121(9), 29.
- Klein, J. (1993b). City of euphemisms. *Newsweek*, 121(8), 33.
- Klein, J. (1993c). Clinton's bushed presidency. *Newsweek*, 122(5), 22.
- Klein, J. (1993d). Clinton's project addiction. *Newsweek*, 121(12), 32.
- Klein, J. (1993e). Clinton's values problem. *Newsweek*, 121(17), 35.
- Klein, J. (1993f, 1993/10/11/). Dinkins's fractured mosaic. *Newsweek*, p. 35.
- Klein, J. (1993g). The education of Berenice Belizaire. (Cover story). *Newsweek*, 122(6), 26.
- Klein, J. (1993h). Elections aren't democracy. *Newsweek*, 122(2), 35.
- Klein, J. (1993i). The end of the chrysanthemums. *Newsweek*, 121(16), 37.
- Klein, J. (1993j). Entering the capital of hell. *Newsweek*, 121(1), 29.
- Klein, J. (1993k, 1993/09/27/). Facing up to the big fear. *Newsweek*, p. 38.
- Klein, J. (1993l). Hail to the chiefs. *Newsweek*, 121(6), 31.
- Klein, J. (1993m). The hidden Lake. *Newsweek*, 122(3), 21.
- Klein, J. (1993n). How about a swift kick? *Newsweek*, 122(4), 30.
- Klein, J. (1993o, 1993/12/06/). Labor's leverage lost. *Newsweek*, p. 25.
- Klein, J. (1993p). 'Make the daddies pay.' *Newsweek*, 121(25), 33.
- Klein, J. (1993q, 1993/09/06/). Michigan's tuna surprise. *Newsweek*, p. 21.
- Klein, J. (1993r). The new neatness. *Newsweek*, 122(26), 21.
- Klein, J. (1993s, 1993/11/15/). New York's rough ride. *Newsweek*, p. 37.
- Klein, J. (1993t). Oops. Maybe we shouldn't have. *Newsweek*, 121(11), 44.
- Klein, J. (1993u). The out-of-wedlock question. *Newsweek*, 122(24), 37.

- Klein, J. (1993v). A poor excuse for poverty solutions. *Newsweek*, 121(3), 27.
- Klein, J. (1993w). Principle or politics? *Newsweek*, 121(24), 29.
- Klein, J. (1993x). Seasick in a rising tide. *Newsweek*, 122(17), 37.
- Klein, J. (1993y). There are jobs in Chicago. *Newsweek*, 122(25), 34.
- Klein, J. (1993z). Time to step back. *Newsweek*, 121(20), 40.
- Klein, J. (1993{). Twilight of the emperor. *Newsweek*, 122(21), 43.
- Klein, J. (1993|). The vice president's ashtray. *Newsweek*, 122(7), 27.
- Klein, J. (1993}, 1993/09/20/). When the Jews could say, 'We won.' (Cover story). *Newsweek*, p. 29.
- Klein, J. (1993~). Window of opportunity. *Newsweek*, 121(10), 35.
- Klein, J. (1994a, 1994/10/31/). Assault on the center. *Newsweek*, p. 33.
- Klein, J. (1994b, 1994/11/14/). An awful year. *Newsweek*, p. 39.
- Klein, J. (1994c, 1994/04/04/). Bill led three lives. *Newsweek*, p. 27.
- Klein, J. (1994d, 1994/08/22/). A bloviational fiesta. *Newsweek*, p. 21.
- Klein, J. (1994e). Bob Reich's job market. *Newsweek*, 123(10), 33.
- Klein, J. (1994f, 1994/09/26/). Bubba is back. *Newsweek*, p. 46.
- Klein, J. (1994g). Can Colin Powell save America? (Cover story). *Newsweek*, 124(15), 20.
- Klein, J. (1994h, 1994/08/01/). Chafee at the bit. *Newsweek*, p. 23.
- Klein, J. (1994i). The citizens of Bimboland. *Newsweek*, 123(1), 59.
- Klein, J. (1994j, 1994/05/30/). Clinton's Caribbean oxymoron. *Newsweek*, p. 54.
- Klein, J. (1994k, 1994/06/20/). The consultants. *Newsweek*, p. 43.
- Klein, J. (1994l). Crime bill garbage barge. *Newsweek*, 123(9), 35.
- Klein, J. (1994m, 1994/10/03/). Empathy for the devil. *Newsweek*, p. 39.

- Klein, J. (1994n, 1994/12/19/). Forever young. *Newsweek*, p. 35.
- Klein, J. (1994o, 1994/06/27/). Grieving for the NAACP. *Newsweek*, p. 37.
- Klein, J. (1994p, 1994/06/06/). 'Hard' vs. 'soft' vs. 'viral' power. *Newsweek*, p. 39.
- Klein, J. (1994q, 1994/05/23/). The health care nose-counters. *Newsweek*, p. 45.
- Klein, J. (1994r). Hi, I'm Joe and I love Tonya too much. *Newsweek*, 123(7), 57.
- Klein, J. (1994s, 1994/04/25/). The House that Newt will build. *Newsweek*, p. 31.
- Klein, J. (1994t, 1994/12/12/). If Chile can do it.....couldn't (North) America privatize its social-security system? *Newsweek*, p. 50.
- Klein, J. (1994u, 1994/06/13/). Learning how to say no. *Newsweek*, p. 29.
- Klein, J. (1994v, 1994/12/05/). Let the states do it. *Newsweek*, p. 35.
- Klein, J. (1994w, 1994/11/07/). Let us now praise famous veggies. *Newsweek*, p. 37.
- Klein, J. (1994x, 1994/07/18/). Plain vanilla or rain forest crunch? *Newsweek*, p. 38.
- Klein, J. (1994y, 1994/07/11/). Primal enterprise. *Newsweek*, p. 37.
- Klein, J. (1994z, 1994/10/24/). A public poet in autumn. *Newsweek*, p. 33.
- Klein, J. (1994{). Puffing motes into dust storms. *Newsweek*, 123(12), 39.
- Klein, J. (1994|, 1994/07/25/). The religious left. *Newsweek*, p. 23.
- Klein, J. (1994}). The Republican eclipse. *Newsweek*, 123(6), 20.
- Klein, J. (1994~, 1994/03/28/). The rites (and wrongs) of spring. *Newsweek*, p. 30.
- Klein, J. (1994□, 1994/08/08/). Robert Kennedy's last campaign. *Newsweek*, p. 23.
- Klein, J. (1994€). Seeing sunshine in Moscow. *Newsweek*, 123(4), 35.
- Klein, J. (1994□, 1994/04/18/). Shepherds of the inner city. *Newsweek*, p. 28.
- Klein, J. (1994,, 1994/09/19/). Smaller, softer, simpler--please. *Newsweek*, p. 34.
- Klein, J. (1994f, 1994/08/15/). A tale of two cities. *Newsweek*, p. 57.

- Klein, J. (1994,,, 1994/03/14/). The threat of tribalism. *Newsweek*, p. 28.
- Klein, J. (1994...). What's German for 'Ross Perot'? *Newsweek*, 123(5), 46.
- Klein, J. (1994†, 1994/11/21/). Wither liberalism? *Newsweek*, p. 56.
- Klein, J. (1995a, 1995/05/15/). AARP? Arrgh. *Newsweek*, p. 27.
- Klein, J. (1995b). Affirmative inaction? *Newsweek*, 125(26), 23.
- Klein, J. (1995c, 1995/04/24/). Back to basic. *Newsweek*, p. 35.
- Klein, J. (1995d, 1995/01/16/). The bear becomes Ursa Minor. *Newsweek*, p. 30.
- Klein, J. (1995e, 1995/03/27/). The birth of common sense. *Newsweek*, p. 31.
- Klein, J. (1995f, 1995/08/07/). The body count. *Newsweek*, p. 34.
- Klein, J. (1995g, 1995/10/02/). Boulevard of broken promises. *Newsweek*, p. 45.
- Klein, J. (1995h, 1995/01/23/). Bowling for virtue. *Newsweek*, p. 26.
- Klein, J. (1995i, 1995/02/20/). Calling Newt's bluff. *Newsweek*, p. 35.
- Klein, J. (1995j, 1995/10/30/). Can Powell reach them? *Newsweek*, p. 48.
- Klein, J. (1995k, 1995/11/13/). Character, not ideology. *Newsweek*, p. 36.
- Klein, J. (1995l, 1995/01/30/). The contemplative bomb-thrower. *Newsweek*, p. 37.
- Klein, J. (1995m). Dances with camels. *Newsweek*, 125(15), 32.
- Klein, J. (1995n, 1995/07/31/). Firm on affirmative action. *Newsweek*, p. 31.
- Klein, J. (1995o, 1995/11/06/). Gingrich and the general. *Newsweek*, p. 40.
- Klein, J. (1995p, 1995/03/20/). How the West was lost. *Newsweek*, p. 31.
- Klein, J. (1995q, 1995/12/11/). Indisputable imponderables. *Newsweek*, p. 35.
- Klein, J. (1995r, 1995/03/13/). Is it dead? *Newsweek*, p. 25.
- Klein, J. (1995s, 1995/10/23/). Lip-syncing the presidency. *Newsweek*, p. 35.
- Klein, J. (1995t, 1995/02/27/). Make a doleful noise. *Newsweek*, p. 80.

- Klein, J. (1995u). Mothers vs. mullahs. *Newsweek*, 125(16), 56.
- Klein, J. (1995v, 1995/05/01/). The nervous '90s. (Cover story). *Newsweek*, p. 58.
- Klein, J. (1995w). Newt 'n the 'hood. *Newsweek*, 126(7), 43.
- Klein, J. (1995x). Off to the culture war. *Newsweek*, 125(24), 28.
- Klein, J. (1995y). The other guy. *Newsweek*, 125(25), 41.
- Klein, J. (1995z, 1995/05/29/). A plausible hothead? *Newsweek*, p. 45.
- Klein, J. (1995{). The Republicans of Madison County. *Newsweek*, 126(9), 41.
- Klein, J. (1995|, 1995/09/18/). Show us your stuff. *Newsweek*, p. 45.
- Klein, J. (1995}, 1995/07/10/). When Bosnia comes home. *Newsweek*, p. 50.
- Klein, J. (1995~). Who are these people? *Newsweek*, 126(8), 31.
- Klein, J. (1995□). Will Newt run for president? *Newsweek*, 125(19), 47.
- Klein, J. (1995€, 1995/10/09/). Will Perot wax the Democrats? *Newsweek*, p. 43.
- Klein, J. (1996a, 1996/04/08/). Advice to the charisma-lorn. *Newsweek*, p. 36.
- Klein, J. (1996b). And now, the age of 'Bibi.' *Newsweek*, 127(24), 34.
- Klein, J. (1996c). Beautiful Bibi. *Newsweek*, 127(22), 40.
- Klein, J. (1996d, 1996/07/29/). A brush with anonymity. *Newsweek*, p. 76.
- Klein, J. (1996e). A case of podium envy. *Newsweek*, 127(24), 33.
- Klein, J. (1996f). Chinese 'face' time. *Newsweek*, 127(23), 46.
- Klein, J. (1996g, 1996/03/04/). Dole finds his mission. *Newsweek*, p. 32.
- Klein, J. (1996h, 1996/08/19/). The end of the tide. *Newsweek*, p. 51.
- Klein, J. (1996i, 1996/07/01/). 'He's come to kick a--.' *Newsweek*, p. 41.
- Klein, J. (1996j, 1996/02/26/). How Clinton could screw up. *Newsweek*, p. 30.
- Klein, J. (1996k). Lame steers in the corral. *Newsweek*, 127(6), 31.

- Klein, J. (1996l, 1996/01/22/). Lawyering the truth. *Newsweek*, p. 34.
- Klein, J. (1996m, 1996/02/12/). Learning to love Dole. *Newsweek*, p. 35.
- Klein, J. (1996n, 1996/09/23/). The limits of negativity. *Newsweek*, p. 42.
- Klein, J. (1996o, 1996/01/29/). Lost in the political fog. *Newsweek*, p. 31.
- Klein, J. (1996p, 1996/09/30/). A lovely donnybrook. *Newsweek*, p. 32.
- Klein, J. (1996q). A lurch toward love. *Newsweek*, 127(8), 37.
- Klein, J. (1996r, 1996/08/12/). Monumental callousness. *Newsweek*, p. 45.
- Klein, J. (1996s, 1996/08/26/). Playing to the squeeze. *Newsweek*, p. 29.
- Klein, J. (1996t, 1996/10/14/). Politics without pols. (Cover story). *Newsweek*, p. 36.
- Klein, J. (1996u, 1996/06/24/). Powell's race problem. *Newsweek*, p. 39.
- Klein, J. (1996v). The predator problem. *Newsweek*, 127(18), 32.
- Klein, J. (1996w, 1996/09/16/). Pretty close to awful. *Newsweek*, p. 551.
- Klein, J. (1996x). The race is to the slow. *Newsweek*, 127(11), 44.
- Klein, J. (1996y). Reports of his death... *Newsweek*, 127(19), 42.
- Klein, J. (1996z, 1996/10/07/). The role of a lifetime. *Newsweek*, p. 43.
- Klein, J. (1996{, 1996/10/21/). Second bananahood. *Newsweek*, p. 42.
- Klein, J. (1996}). Second-term fantasies. *Newsweek*, 127(21), 35.
- Klein, J. (1996}). Severity with a smile. *Newsweek*, 127(2), 41.
- Klein, J. (1996~, 1996/05/27/). Stuck on the periphery. *Newsweek*, p. 39.
- Klein, J. (1996□). The Unabomber and the left. *Newsweek*, 127(17), 39.
- Klein, J. (1996€, 1996/10/28/). Up close and personal. *Newsweek*, p. 40.
- Klein, J. (1996□, 1996/09/09/). Virtual inspiration. *Newsweek*, p. 45.
- Klein, J. (1996,, 1996/11/04/). Where the anger went. *Newsweek*, p. 33.

- Klein, J. (1996*f*, 1996/01/15/). The year of the nerd? *Newsweek*, p. 37.
- Kretzschmar, W., Meyer, C., & Ingegneri, D. (1997). Uses of inferential Statistics in Corpus Studies. In M. Ljung (Ed.), *Corpus-based Studies in English* (pp. 167-177). Amsterdam: Rodopi.
- Love, H. (2002). *Attributing authorship: an introduction*. New York: Cambridge University Press.
- Maloney, J. J., & O'Connor, J. P. (1999, May 7, 1999). The Murder of JonBonet Ramsey. *Crime Magazine*, 1-45.
- McMenamin, G. R. (2001). Style Markers in Authorship Studies. *Forensic Linguistics*, 8(2), 93-97.
- McMenamin, G. R. (2002). *Forensic Linguistics: Advances in Forensic Stylistics*. New York: CRC Press.
- Melton, J. G. (1998). *Finding Enlightenment: Ramtha's School of Ancient Wisdom*. Hillsboro: Beyond Words Publishing, Inc.
- Mendenhall, T. C. (1887). The characteristic Curves of Composition. *Science*, 11, 237-249.
- Montgomery, R. S. (1971). *A World Beyond*. New York: Fawcett.
- Montgomery, R. S. (1976). *The World Before*. New York: Coward McCann & Geoghegan.
- Montgomery, R. S. (1999). *The World to Come*. New York: Harmony Books.
- Mosteller, F. (1987). A Statistical Study of the Writing Styles of Authors of the Federalist Papers. *Proceedings of the American Philosophical Society*, 131(2), 132-140.
- NearDeathExperiences. (2005). *Ruth Montgomery, Psychic about the afterlife*. Retrieved October 15, 2005, from

- <http://paranormal.about.com/gi/dynamic/offsite.htm?zi=1/XJ&sdn=paranormal&zu=http://www.near-death.com/experiences/paranormal10.html>
- Norušis, M. J. (2005). *SPSS 13.0 Statistical Procedures Companion*. Upper Saddle River: Prentiss Hall, Inc.
- Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Olsson, J. (2004). *Forensic Linguistics: An Introduction to Language, Crime and the Law*. London: Continuum.
- Radin, D. (2003). Some evidence indicates that Psychic ability exists. In *Opposing View Points: Paranormal Phenomena*. San Diego: Greenhaven Press.
- Ramtha, K., J.Z.). (1985 - current). *Mini-teachings*: retrieved from www.ramtha.com/html/community/teachings/default.stm.
- Rice, A. (2004). from www.annerice.com
- Rudman, J. (2002). *Non-Traditional Authorship Attribution Studies in Eighteenth Century Literature. Stylistics Statistics and the Computer*. Retrieved November 1, 2005, from <http://computerphilologie.uni-muenchen.de/jg02/rudman.html>
- Scott, M. (2004). *WordSmith Tools 4.0*. Retrieved December 27 2005, 2006, from <http://www.lexically.net/wordsmith/index.html>
- Shermer, M. (1997). *Why people believe weird things*. New York: Henry Holt.
- Shermer, M. (2005). Annual expenditures - Parapsychology Industry. Personal Correspondence: July 19/20 2005.
- Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Svartvik, J. (1968). *The Evans Statements; a case for forensic linguistics*. (Vol. 20): Goteborg, Elanders boktryckeri aktiebolag.

TimeMagazine. (2006). *Joe Klein Biography*, from

<http://www.time.com/time/columnist/klein/article/0,9565,490843,00.html>

Tribbe, F. C. (Ed.). (1996). *An Arthur Ford Anthology*. Nevada City: Blue Dolphin Publishing, Inc.

Tweedie, F. J., Holmes, D. I., & Corns, T. (1998). The Provenance of De Doctrina Christiana, attributed to John Milton: A Statistical Investigation. *Literary and Linguistic Computing*, 13(2), 77-87.

UCREL. (2005). Retrieved July 11, 2005, from

<http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/trial.html>

WashingtonPost. (1996, February 02, 1996). Wanted Anonymous: Sure, They Deny It. But If They Didn't Do It, Who Did. *The Washington Post*, p. B01.

Wikipedia. (2005a). *Ramtha*. Retrieved October 25, 2005, from

http://en.wikipedia.org/wiki/Main_Page

Wikipedia. (2005b). *Sylvia Browne*. Retrieved October 11, 2005, from

http://en.wikipedia.org/wiki/Main_Page

Yule, G. U. (1939). On sentence-length as a statistical characteristic of style in prose: with application to two cases of disputed authorship. *Biometrika*, 30(3/4), 363-390.

Appendix A - Psychic Matrix Raw Data

Rank			1	2	3	4	5
Author	Texts	Segment	THE	OF	AND	TO	THAT
Montgomery	1 WBM	1	241	122	137	149	83
Montgomery	1 WBM	2	239	148	127	151	76
Montgomery	1 WBM	3	134	88	133	78	53
Ford	2 WBF	1	225	140	165	176	89
Ford	2 WBF	2	218	104	140	208	106
Ford	2 WBF	3	284	194	135	190	84
Ford	2 WBF	4	269	177	184	155	93
Ford	2 WBF	5	212	110	157	182	107
Ford	2 WBF	6	207	129	173	191	73
Ford	2 WBF	7	238	155	145	189	76
Ford	2 WBF	8	242	153	145	177	74
Ford	2 WBF	9	265	178	171	187	72
Ford	3 WWSFORD	1	256	230	139	112	16
Ford	3 WWSFORD	2	364	206	161	83	103
Ford	3 WWSFORD	3	310	197	163	106	14
Ford	3 WWSFORD	4	293	182	137	107	105
Ford	3 WWSFORD	5	364	239	154	124	83
Ford	4 UBKFord	1	290	217	129	130	70
Ford	4 UBKFord	2	253	213	125	121	52
Ford	4 UBKFord	3	285	208	143	120	23
Ford	4 UBKFord	4	303	183	110	123	57
Ford	4 UBKFord	5	229	174	107	111	69
Ford	4 UBKFord	6	210	159	109	154	59
Ford	4 UBKFord	7	261	178	128	113	93
Ford	4 UBKFord	8	310	169	188	125	82
Ford	4 UBKFord	9	282	166	122	163	63
Ford	4 UBKFord	10	258	107	103	117	58
Montgomery	5 WTCMONT	1	234	110	184	148	90
Montgomery	5 WTCMONT	2	323	160	174	167	92
Montgomery	5 WTCMONT	3	327	150	170	141	100
Montgomery	5 WTCMONT	4	298	146	163	167	94
Montgomery	5 WTCMONT	5	258	140	165	151	114
Montgomery	5 WTCMONT	6	286	140	193	170	101
Montgomery	6 TWB	1	337	204	175	125	74
Montgomery	6 TWB	2	311	189	192	123	82
Montgomery	6 TWB	3	368	196	165	131	68
Montgomery	6 TWB	4	326	176	160	142	83
Montgomery	6 TWB	5	291	154	184	188	85
Montgomery	6 TWB	6	404	214	153	108	78
Montgomery	6 TWB	7	415	217	197	151	80
Montgomery	6 TWB	8	351	211	185	115	82
Montgomery	6 TWB	9	314	217	192	154	66

6	7	8	9	10	11	12	13	14
A	IN	IS	YOU	IT	I	ARE	AS	WAS
107	94	26	50	26	94	16	39	66
97	119	36	25	22	124	33	36	37
64	69	35	11	16	34	41	32	12
79	60	113	45	58	54	65	95	33
94	83	90	22	32	18	64	50	24
93	93	98	19	60	33	74	39	8
94	94	91	21	24	1	60	63	29
82	109	63	16	18	38	36	47	26
99	90	105	26	25	10	59	54	24
93	105	78	13	36	2	64	33	30
97	88	81	58	47	50	50	71	34
110	101	75	25	47	14	81	49	11
153	111	121	9	65	47	60	47	11
115	113	111	17	62	28	48	45	9
113	118	79	22	50	48	51	79	15
110	119	82	3	77	134	30	51	37
78	106	102	8	50	63	21	42	36
109	95	49	2	45	79	10	28	79
142	124	29	5	26	28	21	38	72
120	107	35	5	37	82	7	26	72
139	109	66	12	46	38	33	36	32
142	80	64	32	45	67	25	31	55
122	94	56	60	45	85	21	24	51
115	87	90	33	52	76	49	42	47
99	96	114	94	53	34	47	21	28
102	94	93	70	76	48	46	43	25
126	102	53	5	45	57	13	31	45
108	75	22	32	36	152	11	33	72
83	77	40	25	56	69	33	46	42
64	85	80	21	59	43	58	64	24
88	97	37	14	37	57	45	54	26
108	108	42	27	39	69	17	74	84
89	96	109	56	74	98	49	52	78
89	96	26	5	26	40	22	60	58
65	108	22	4	21	18	16	56	57
71	95	16	6	35	25	10	38	58
90	109	33	18	41	24	19	47	43
86	104	17	56	38	27	14	38	114
70	100	13	3	29	10	17	43	57
78	113	20	1	26	21	8	44	65
76	107	44	4	48	8	27	30	42
69	142	35	2	31	8	31	30	67

15	16	17	18	19	20	21	22	23
FOR	THEY	THIS	BE	HAVE	WITH	WE	NOT	BUT
39	12	27	18	10	46	25	21	33
53	13	26	16	17	32	38	20	22
39	27	19	16	27	17	27	28	23
79	14	56	27	15	33	0	37	38
73	53	44	24	27	36	43	32	41
68	30	458	32	25	31	59	28	43
64	40	44	24	25	35	60	39	27
65	47	36	18	24	49	29	30	36
61	47	44	34	27	47	24	37	43
73	49	41	20	16	35	62	39	53
76	28	41	31	31	31	51	34	35
64	34	53	55	29	38	72	48	50
27	43	26	41	20	26	58	58	38
18	19	39	26	29	40	56	52	39
32	40	30	61	42	43	54	52	36
23	41	48	37	40	27	24	52	26
31	21	36	35	25	17	20	29	21
28	16	32	24	27	46	6	23	10
25	14	34	23	19	31	8	17	9
39	14	24	24	16	36	8	21	16
26	14	25	28	40	45	19	26	13
28	13	43	19	44	52	25	33	19
27	12	33	38	31	30	26	36	31
19	39	20	31	32	41	24	26	40
28	26	23	52	36	23	11	38	50
31	17	30	45	42	33	18	31	26
38	12	27	25	22	33	8	29	15
45	18	14	22	27	50	32	29	32
50	43	16	48	32	36	42	36	30
39	68	25	69	22	28	50	44	43
53	71	20	85	38	28	20	33	44
45	42	29	40	11	30	11	21	29
39	45	34	41	49	30	43	37	51
50	33	39	13	33	52	27	22	27
44	75	17	11	19	45	14	17	22
50	56	20	11	9	39	9	17	21
43	49	20	22	27	36	7	19	15
57	31	33	14	14	39	14	28	31
48	36	26	8	17	37	12	22	21
49	42	29	20	15	45	4	18	19
39	55	25	16	29	51	21	20	29
38	35	27	16	24	38	27	26	29

24	25	26	27	28	29	30	31	32
HE	WILL	ON	WHO	YOUR	ALL	SO	WERE	FROM
120	9	41	30	0	8	17	16	21
86	13	44	32	4	36	19	10	19
20	9	30	11	5	12	11	4	12
37	28	39	17	16	40	36	9	17
27	52	18	39	7	14	33	8	6
52	21	30	36	1	28	37	11	16
82	29	28	33	5	27	24	10	18
176	27	29	37	6	13	21	17	15
100	42	39	41	5	14	26	7	12
104	38	31	40	5	13	33	10	15
50	22	39	34	14	21	31	15	26
13	86	35	33	9	17	32	15	25
27	8	13	18	9	30	13	8	15
35	7	18	11	17	37	22	3	17
36	48	21	25	7	26	17	6	14
28	17	22	27	8	29	9	16	14
27	18	10	14	0	26	10	1	26
41	8	22	18	1	16	23	21	21
51	10	32	19	0	14	15	21	29
41	6	26	16	2	13	12	26	16
17	4	37	22	2	13	11	17	27
36	2	28	29	10	14	15	9	21
88	19	37	23	13	16	16	10	23
78	9	29	22	9	13	15	7	22
81	49	19	54	15	18	13	6	17
42	32	19	29	27	26	10	5	15
28	5	25	21	2	10	13	10	17
62	7	35	27	11	24	17	14	19
15	55	30	29	6	27	21	12	15
18	110	24	24	2	32	20	9	22
24	132	33	33	4	13	20	14	16
60	75	33	34	5	29	12	16	14
92	50	22	42	18	29	38	15	19
13	19	32	27	4	15	18	37	17
15	11	32	25	4	12	25	47	34
12	4	39	33	1	15	20	78	45
8	3	31	27	4	13	25	54	39
25	2	28	28	28	22	31	53	36
7	5	36	24	1	12	16	57	36
18	12	22	30	0	11	17	38	39
7	13	18	40	2	9	22	38	41
40	23	21	53	0	16	32	41	22

33	34	35	36	37	38	39	40	41
BY	WHAT	THERE	HAD	OR	THEIR	ONE	HIS	MY
21	10	8	49	11	3	10	63	36
17	12	8	46	11	5	24	39	34
7	8	14	10	28	5	15	17	9
16	31	40	18	37	9	22	17	17
15	13	27	7	27	17	20	54	12
20	15	22	16	25	11	16	25	6
26	8	22	21	14	23	21	47	0
18	13	13	52	21	19	15	72	10
9	7	15	21	19	24	18	42	3
18	12	21	19	33	13	13	48	0
21	10	21	19	36	13	19	29	20
24	15	14	7	40	21	14	10	3
20	16	17	9	30	11	13	24	14
21	13	17	6	33	7	25	26	14
36	9	34	13	16	19	26	20	16
28	11	22	47	27	9	21	11	28
21	20	16	13	39	9	14	18	18
29	16	7	41	11	12	22	22	43
31	15	8	27	26	10	32	55	17
40	5	10	52	15	15	18	36	29
40	8	20	23	27	8	25	33	15
28	11	17	42	18	7	22	34	40
14	10	20	26	18	6	22	41	19
28	14	35	20	31	8	24	20	11
15	18	41	9	20	6	37	27	3
21	9	25	5	27	7	31	21	15
37	6	8	35	17	12	30	36	33
13	11	20	49	11	5	12	35	69
20	14	24	27	25	25	11	11	25
22	17	22	15	15	27	14	11	12
30	13	18	25	15	34	10	19	20
22	6	8	47	5	17	14	28	30
18	28	35	32	12	9	23	49	13
33	5	19	21	22	20	6	17	23
30	8	19	35	15	38	24	24	2
29	5	18	34	23	34	17	19	9
41	1	27	30	28	32	18	8	4
23	2	21	31	14	19	21	22	19
38	10	17	42	18	34	16	6	0
31	3	25	40	12	28	16	6	6
23	9	15	28	17	39	10	10	2
26	3	20	28	21	25	14	25	0

42	43	44	45	46	47	48	49	50	
WHEN	AN	AT	IF	WHICH	INTO	HAS	CAN	BECAUSE	Segment Length
15	15	31	7	22	6	8	8	7	5041
19	12	17	16	23	11	11	4	18	5030
15	11	13	12	15	5	8	8	3	2992
12	6	22	15	32	8	10	9	13	5033
16	8	22	21	20	13	30	7	14	5012
11	8	30	19	18	16	10	1	16	5023
16	147	23	14	26	16	22	1	8	5027
20	13	20	6	17	19	23	1	13	5017
16	12	24	3	15	11	20	3	13	5018
16	16	11	9	33	14	21	5	19	5034
15	21	15	12	29	16	12	1	17	5019
21	8	17	13	25	19	7	2	11	5336
15	24	9	14	30	14	31	31	9	4990
17	13	20	8	32	15	28	19	4	4998
9	13	14	24	32	25	12	13	5	4993
13	17	22	23	25	28	14	13	4	4992
14	15	12	13	28	13	11	8	5	4739
18	15	23	7	18	12	13	2	5	5031
19	21	29	3	10	11	12	6	3	5048
10	14	35	8	9	8	11	6	2	5068
18	20	30	11	11	10	17	8	5	5048
19	22	36	10	14	8	18	4	4	5055
14	26	22	14	7	1	12	9	2	5034
14	13	23	10	21	8	21	11	9	5016
20	9	9	9	43	11	45	26	24	5006
17	17	12	28	9	20	21	19	10	5007
16	22	16	12	17	9	26	10	3	4854
13	16	26	14	11	9	6	3	6	5015
15	14	19	10	4	14	5	14	8	5021
10	18	21	7	3	16	8	10	11	5011
14	6	14	9	5	8	9	4	4	5028
17	7	19	7	5	8	8	5	8	5004
10	14	21	23	4	20	22	24	17	5749
11	14	20	5	14	20	5	5	8	5040
12	18	17	11	23	12	8	3	3	5022
21	9	17	7	18	11	4	1	10	5031
19	12	19	7	21	8	8	3	4	5022
18	9	17	7	16	12	10	1	8	5026
16	14	20	4	19	5	4	1	4	5012
8	13	19	9	19	8	3	1	5	5041
14	11	15	5	15	13	7	1	6	5015
21	21	16	2	20	12	6	2	12	5026

Rank			1	2	3	4	5
Author	Texts	Segment	THE	OF	AND	TO	THAT
Montgomery	6 TWB	10	292	211	186	147	71
Montgomery	6 TWB	11	201	165	145	192	82
Montgomery	6 TWB	12	265	162	132	153	74
Montgomery	6 TWB	13	298	170	194	150	73
Montgomery	6 TWB	14	138	83	75	97	34
Brown	7 CFOSBROWN	1	164	115	96	160	88
Brown	7 CFOSBROWN	2	219	103	122	120	66
Brown	7 CFOSBROWN	3	242	156	158	135	55
Brown	7 CFOSBROWN	4	215	146	174	131	56
Brown	7 CFOSBROWN	5	219	167	124	158	66
Brown	7 CFOSBROWN	6	323	144	127	140	58
Ramtha (Knight, JZ)	8 RAMTHA	1	268	142	226	115	129
Ramtha (Knight, JZ)	8 RAMTHA	2	257	161	195	157	183
Ramtha (Knight, JZ)	8 RAMTHA	3	234	123	177	137	149
Ramtha (Knight, JZ)	8 RAMTHA	4	161	146	211	151	161
Ramtha (Knight, JZ)	8 RAMTHA	5	289	125	230	139	161
Ramtha (Knight, JZ)	8 RAMTHA	6	248	101	181	143	164
Ramtha (Knight, JZ)	8 RAMTHA	7	208	118	189	171	196
Ramtha (Knight, JZ)	8 RAMTHA	8	290	139	188	138	159
Ramtha (Knight, JZ)	8 RAMTHA	9	234	132	188	154	200
Ramtha (Knight, JZ)	8 RAMTHA	10	299	214	153	169	138
Ramtha (Knight, JZ)	8 RAMTHA	11	261	111	220	130	141
Ramtha (Knight, JZ)	8 RAMTHA	12	209	89	223	145	191
Ramtha (Knight, JZ)	8 RAMTHA	13	242	144	169	160	171

6	7	8	9	10	11	12	13	14
A	IN	IS	YOU	IT	I	ARE	AS	WAS
104	153	34	7	26	19	37	49	82
105	101	70	24	87	90	32	57	59
107	111	83	17	48	57	55	73	86
100	105	82	8	47	4	55	46	3
65	47	49	10	28	1	44	44	4
108	61	84	245	46	43	38	35	4
87	85	67	239	51	42	30	30	4
111	76	78	155	33	24	43	40	0
98	84	89	44	35	11	41	40	4
117	90	69	83	60	28	42	50	4
103	86	66	77	44	8	22	33	9
98	91	119	170	94	16	57	24	44
73	66	195	211	131	21	68	15	24
120	99	166	186	95	19	77	19	31
81	105	136	140	96	37	57	10	23
81	89	106	92	132	133	46	81	59
81	81	181	256	153	49	150	12	14
97	83	139	193	135	120	82	35	29
79	55	132	181	131	80	41	59	50
88	62	190	250	132	54	61	32	17
63	75	146	292	83	13	96	15	5
90	65	103	115	122	90	32	35	91
58	83	132	194	162	190	52	72	69
82	66	230	295	170	30	96	17	11

15	16	17	18	19	20	21	22	23
FOR	THEY	THIS	BE	HAVE	WITH	WE	NOT	BUT
53	48	39	24	24	36	40	21	22
72	24	39	20	26	35	59	35	39
65	27	46	37	26	27	63	21	35
65	28	38	66	27	35	34	36	39
24	25	18	31	18	18	31	24	21
40	23	54	47	40	35	14	21	41
36	18	67	46	31	30	16	19	27
52	18	51	35	50	28	35	18	25
59	23	53	31	45	36	79	12	18
58	39	46	29	42	29	38	16	25
58	55	47	33	26	23	23	17	24
32	91	23	48	40	19	2	39	14
30	54	27	35	51	20	3	35	27
23	48	26	32	60	22	94	33	22
17	44	47	25	50	26	83	17	29
20	31	23	32	23	16	28	30	38
10	46	73	36	43	20	92	28	20
19	53	65	26	79	18	42	41	20
28	23	34	32	31	12	42	24	29
22	22	70	34	54	20	32	18	25
61	19	34	40	60	14	4	22	16
37	34	38	29	29	28	16	37	29
18	18	28	38	33	15	2	35	42
33	25	47	34	66	21	19	40	27

24	25	26	27	28	29	30	31	32
HE	WILL	ON	WHO	YOUR	ALL	SO	WERE	FROM
32	12	35	32	0	15	26	29	23
29	18	30	24	2	21	35	23	15
13	59	30	23	1	21	24	16	17
23	176	27	40	0	17	26	5	26
6	41	10	15	1	9	11	3	13
3	27	22	22	82	19	21	8	19
5	27	21	7	99	15	24	4	13
1	13	72	13	85	31	24	1	9
9	9	88	12	47	46	16	4	16
2	24	46	29	40	28	19	2	26
3	24	30	29	32	41	21	7	17
20	17	18	24	53	19	29	41	20
12	11	26	33	46	31	21	30	4
10	14	19	36	51	31	43	13	5
55	7	18	46	27	20	36	23	21
13	16	11	12	12	23	20	61	25
0	15	12	8	36	34	19	6	27
2	13	23	12	18	21	27	30	24
51	66	21	23	72	20	25	35	27
11	15	19	14	47	28	48	20	14
6	38	8	26	101	42	14	9	14
82	6	13	28	39	31	48	33	21
5	16	6	9	54	32	27	49	15
6	25	28	22	77	32	35	16	11

33	34	35	36	37	38	39	40	41
BY	WHAT	THERE	HAD	OR	THEIR	ONE	HIS	MY
30	8	22	27	15	30	14	22	8
18	15	25	36	18	11	32	23	20
12	15	31	14	22	19	21	25	11
21	4	25	1	17	17	14	13	0
15	10	4	2	16	13	6	6	0
22	16	13	10	29	19	24	5	9
16	13	5	2	40	11	18	0	9
14	7	11	1	36	10	15	1	9
15	9	16	1	53	19	15	10	10
16	12	13	4	36	35	16	2	16
15	18	8	1	35	53	14	4	2
10	29	36	10	17	51	17	18	5
13	37	28	13	7	18	14	3	9
9	31	15	10	17	18	13	18	11
15	48	20	6	8	25	20	27	4
8	45	24	14	1	18	13	14	46
7	68	26	2	7	15	18	1	3
9	42	33	10	10	27	15	3	12
9	44	16	10	12	11	11	22	13
7	74	24	5	15	8	8	5	16
9	38	15	1	12	6	14	3	2
5	54	31	25	10	8	22	18	67
7	61	19	17	2	5	17	2	83
6	77	18	5	10	10	10	1	8

42	43	44	45	46	47	48	49	50	
WHEN	AN	AT	IF	WHICH	INTO	HAS	CAN	BECAUSE	Segment Length
24	16	11	12	7	8	8	2	7	5020
14	17	22	22	22	4	7	3	10	5004
18	16	25	16	21	14	9	2	11	5004
20	8	11	11	15	12	14	3	7	5010
15	6	6	8	6	5	3	4	6	2541
12	19	7	49	7	10	16	27	16	5021
24	22	10	42	5	20	9	34	23	5012
19	20	17	18	9	15	12	33	10	5034
21	28	8	24	8	10	19	29	22	5029
27	28	9	20	5	10	21	30	11	5027
21	45	13	20	10	17	9	37	14	4999
23	5	19	13	20	29	16	10	19	5035
16	9	15	27	39	13	20	11	21	5044
24	11	11	19	7	18	12	25	40	5062
12	14	4	12	17	12	20	27	24	5054
18	12	13	19	28	24	8	8	17	5026
22	10	13	25	7	14	13	50	42	5078
26	23	16	16	4	20	11	26	32	5037
23	19	12	15	9	16	13	14	14	5036
15	24	13	36	8	13	13	14	36	5039
17	13	5	12	16	35	19	15	25	5036
16	14	11	20	19	20	17	2	17	5025
21	10	10	25	22	20	5	20	44	5032
16	20	9	47	30	17	16	12	31	5215

Appendix B - Genre Matrix Raw Data

Rank			1	2	3	4	5
Author	Texts	Segment	the	of	to	a	and
Anonymous	Primary Colors	1	204	90	90	150	127
Anonymous	Primary Colors	2	227	94	91	138	83
Anonymous	Primary Colors	3	195	78	102	122	102
Anonymous	Primary Colors	4	224	56	114	137	81
Anonymous	Primary Colors	5	235	84	116	114	119
Anonymous	Primary Colors	6	250	61	116	111	114
Anonymous	Primary Colors	7	206	73	109	108	117
Anonymous	Primary Colors	8	215	67	130	81	89
Anonymous	Primary Colors	9	242	96	104	113	109
Anonymous	Primary Colors	10	177	70	89	133	91
Anonymous	Primary Colors	11	189	76	131	110	132
Anonymous	Primary Colors	12	201	59	117	110	98
Anonymous	Primary Colors	13	199	78	114	111	98
Anonymous	Primary Colors	14	207	70	99	141	101
Anonymous	Primary Colors	15	218	69	105	112	137
Anonymous	Primary Colors	16	223	74	88	132	123
Anonymous	Primary Colors	17	201	90	111	126	104
Anonymous	Primary Colors	18	181	68	108	114	116
Anonymous	Primary Colors	19	189	80	100	125	96
Anonymous	Primary Colors	20	192	86	115	109	123
Anonymous	Primary Colors	21	205	79	99	114	104
Anonymous	Primary Colors	22	204	87	111	140	103
Anonymous	Primary Colors	23	200	77	115	117	114
Anonymous	Primary Colors	24	222	87	109	116	135
Anonymous	Primary Colors	25	185	61	89	143	150
Anonymous	Primary Colors	26	169	77	95	111	125
Anonymous	Primary Colors	27	218	94	113	119	130
Anonymous	Primary Colors	28	178	77	128	91	91
Anonymous	Primary Colors	29	55	19	46	20	31
Joe Klein	Public Lives	1	294	106	120	136	111
Joe Klein	Public Lives	2	242	110	101	146	123
Joe Klein	Public Lives	3	318	127	134	125	113
Joe Klein	Public Lives	4	267	104	127	146	108
Joe Klein	Public Lives	5	290	127	109	147	106
Joe Klein	Public Lives	6	257	111	133	136	131
Joe Klein	Public Lives	7	285	106	112	126	117
Joe Klein	Public Lives	8	300	111	157	125	128
Joe Klein	Public Lives	9	291	126	128	133	102
Joe Klein	Public Lives	10	319	133	143	133	116
Joe Klein	Public Lives	11	250	117	126	122	108
Joe Klein	Public Lives	12	285	111	157	140	103
Joe Klein	Public Lives	13	250	108	134	121	126
Joe Klein	Public Lives	14	308	118	129	130	98
Joe Klein	Public Lives	15	320	116	128	129	106
Joe Klein	Public Lives	16	341	135	114	123	108

6	7	8	9	10	11	12	13	14
in	is	that	he	it	not	for	was	on
63	32	42	116	56	64	28	96	30
64	45	37	111	94	75	29	130	36
79	53	33	161	90	66	30	97	36
70	37	47	131	68	55	38	95	45
57	41	54	104	91	64	37	103	39
60	43	30	71	87	57	34	101	49
58	47	43	99	90	80	22	107	38
56	59	63	62	95	68	31	75	33
52	35	43	68	70	61	45	113	39
59	54	51	62	101	77	29	106	34
61	34	41	92	64	70	36	96	48
50	46	48	97	96	75	36	92	24
77	60	44	68	62	63	37	58	39
65	45	43	93	60	42	48	70	39
70	19	48	74	82	54	25	114	58
79	62	39	103	64	64	32	72	32
78	66	52	109	83	74	32	65	38
84	39	40	100	96	66	35	96	52
70	61	54	66	86	90	33	59	32
63	61	56	85	62	63	43	74	20
58	39	42	125	76	83	32	98	28
71	50	59	87	62	70	25	76	39
60	40	49	74	85	96	26	76	37
60	55	44	83	78	52	45	74	35
67	38	36	97	69	63	32	102	32
49	59	38	58	92	67	24	68	41
77	32	48	31	92	56	42	98	28
53	30	58	97	114	88	36	120	34
11	21	7	33	23	19	13	18	8
93	87	55	116	67	55	35	76	24
75	100	62	83	63	55	33	18	22
57	63	60	49	68	56	48	76	44
74	91	74	75	54	63	44	40	30
76	85	44	84	49	49	42	60	22
88	90	63	101	90	46	41	30	21
94	113	41	96	59	47	40	35	26
57	98	76	76	56	64	42	31	32
82	103	40	112	59	64	35	21	22
73	84	45	55	51	50	42	38	40
89	94	52	107	60	62	51	19	30
82	52	42	50	40	49	45	50	27
69	82	47	61	48	63	49	20	22
102	69	60	41	63	42	40	44	26
91	94	46	78	58	51	45	25	42
77	74	41	39	77	52	48	44	33

15	16	17	18	19	20	21	22	23
his	as	but	with	be	are	by	this	has
43	28	25	32	19	22	7	20	3
33	15	32	28	22	20	7	29	0
35	11	42	32	19	31	7	18	2
36	15	23	43	23	17	5	27	2
26	24	28	30	32	16	10	31	3
22	10	35	28	25	36	4	22	2
17	11	32	26	22	47	7	41	2
24	18	33	38	20	28	4	37	7
22	23	45	32	45	18	15	43	3
14	9	35	27	22	29	9	17	2
35	26	28	26	34	36	10	35	5
17	12	32	25	36	26	12	37	1
15	21	31	26	34	45	10	36	3
28	28	34	33	33	27	10	39	10
16	9	37	29	14	21	11	32	0
36	17	27	21	25	29	10	43	9
27	25	48	43	28	23	10	29	10
42	15	37	37	15	20	8	40	2
34	19	33	38	20	23	13	29	3
23	15	35	31	40	34	6	47	9
19	10	38	33	28	21	15	25	3
18	19	34	36	22	21	12	30	5
16	15	46	30	28	30	8	34	1
24	14	39	26	29	31	6	34	6
32	23	40	27	25	23	7	32	3
21	15	29	38	25	45	13	49	11
16	17	32	40	25	22	9	22	4
49	17	49	17	28	18	8	33	2
16	4	10	8	8	20	2	17	0
55	29	45	25	28	12	17	28	23
42	38	38	29	42	30	23	26	37
23	23	55	29	43	15	23	22	16
32	31	38	27	40	28	27	28	33
35	30	39	31	35	32	24	19	21
50	26	43	28	24	28	21	18	28
43	28	39	20	29	35	18	14	24
24	30	48	26	51	37	15	31	37
42	26	46	22	32	43	17	17	29
29	38	32	25	46	32	13	27	31
51	21	30	28	50	33	17	30	40
31	29	37	32	38	30	21	24	14
27	27	42	33	46	39	21	25	28
24	34	36	22	40	27	14	28	38
21	17	40	29	24	32	18	39	45
17	34	39	25	43	24	22	23	21

24	25	26	27	28	29	30	31	32
have	I	they	who	had	at	an	would	will
14	117	28	16	41	20	16	17	6
31	135	31	5	28	13	11	39	13
12	114	17	15	31	22	9	33	17
26	87	17	14	29	22	10	25	17
24	58	16	12	45	18	13	36	8
17	115	23	12	21	25	12	22	14
17	146	38	17	31	28	12	14	8
20	136	17	17	45	20	9	34	15
21	84	37	20	50	18	12	53	18
28	136	30	11	21	39	19	24	22
31	96	15	10	53	21	17	24	19
23	138	23	22	29	50	16	32	25
35	136	20	18	28	19	13	26	15
29	106	24	12	21	21	12	38	15
21	139	13	3	37	19	16	15	17
11	107	15	12	14	25	12	23	19
26	124	26	13	25	19	13	21	19
22	117	4	12	27	40	13	17	11
37	135	22	11	23	21	17	27	27
42	200	19	5	28	26	8	15	30
20	147	21	13	24	24	5	18	9
18	133	17	5	35	27	20	21	18
39	188	23	8	37	23	7	19	20
28	110	13	12	13	20	15	26	19
16	157	12	11	18	27	10	18	19
25	127	24	13	33	33	11	21	15
33	225	13	8	26	20	22	30	15
35	261	37	12	31	22	10	36	15
12	35	6	3	4	5	3	7	1
23	24	12	17	18	19	32	15	13
27	21	23	13	10	20	31	21	19
30	68	19	27	24	12	17	28	28
27	24	10	15	16	15	20	22	32
16	25	17	20	21	11	28	22	20
25	20	13	23	14	17	15	20	26
22	32	32	23	11	16	16	15	18
37	25	25	18	17	24	29	37	29
33	36	45	22	6	15	14	21	21
26	13	32	16	14	14	14	12	29
28	36	20	23	7	16	18	8	50
37	30	18	17	16	12	24	13	25
33	14	40	19	5	16	18	19	22
37	18	27	15	11	17	21	31	15
30	14	27	21	11	13	19	18	17
30	5	19	17	10	18	18	31	25

33	34	35	36	37	38	39	40	41
there	from	said	president	about	more	one	we	or
26	24	40	3	20	6	17	22	16
33	9	59	2	21	8	17	40	18
17	16	60	1	33	14	10	34	22
18	14	54	3	18	7	13	60	6
16	22	41	2	37	9	12	54	13
15	14	55	2	20	4	12	49	10
24	18	67	2	30	3	14	85	8
17	18	92	1	45	6	15	76	5
24	17	41	3	34	7	17	75	13
31	19	58	1	18	3	11	54	15
30	13	48	2	30	10	11	101	9
21	14	56	1	15	5	11	75	24
26	17	60	2	16	7	11	62	13
27	12	64	3	27	6	12	46	8
26	15	64	2	15	13	6	67	13
25	12	66	2	34	10	13	52	9
26	14	48	4	23	12	15	36	12
15	7	80	1	23	8	13	65	12
16	17	68	5	20	11	19	35	20
25	11	78	1	34	8	13	51	16
23	18	68	1	15	8	7	37	13
31	8	69	3	23	10	10	24	17
21	17	66	4	29	8	13	60	10
31	15	72	3	15	11	11	69	12
25	17	74	3	14	5	4	46	7
17	17	86	0	16	5	19	58	15
33	16	53	1	25	3	9	28	6
23	13	52	4	30	8	12	14	4
13	5	15	3	16	2	4	22	1
28	18	13	27	30	17	14	4	7
14	11	9	17	28	15	11	36	7
25	16	22	10	21	14	12	13	15
16	15	7	23	18	19	15	8	16
14	16	9	5	15	15	15	9	15
24	12	13	17	27	17	13	17	6
20	8	25	10	26	15	20	11	12
34	16	10	18	15	14	8	16	7
14	7	12	12	23	28	8	18	8
24	13	17	14	27	13	20	10	14
18	10	18	18	30	31	19	18	12
15	12	19	9	15	18	13	26	8
20	13	11	7	15	22	1	30	13
27	16	8	12	12	25	16	11	13
25	24	10	22	22	17	17	10	9
30	14	9	19	7	17	13	13	16

42	43	44	45	46	47	48	49	50	Segment
were	do	been	you	all	what	if	their	so	Length
24	31	11	85	19	27	18	7	27	5042
18	26	5	91	16	22	11	8	15	4993
22	18	4	104	14	23	21	7	16	5070
12	21	9	114	16	13	22	5	27	5015
31	29	9	74	18	26	17	5	9	5042
20	31	5	71	15	14	3	7	19	5089
28	41	9	99	21	17	6	6	12	5025
19	41	8	71	23	43	12	2	17	5063
29	36	16	56	20	24	28	4	10	5068
18	32	9	101	13	24	9	6	22	5015
36	26	22	75	21	14	20	8	19	5025
20	29	8	82	20	27	15	6	26	5044
17	52	10	121	17	39	25	5	26	5077
23	21	18	95	29	19	21	7	10	5114
21	22	10	60	20	18	8	3	15	5067
8	36	7	57	15	20	8	3	15	5076
17	33	7	65	13	25	21	4	11	5077
23	27	13	70	11	35	8	9	17	5066
8	35	12	103	21	27	29	4	23	5062
14	40	12	98	19	33	15	5	21	5075
26	31	13	94	12	25	17	5	15	5071
22	38	16	79	18	19	14	3	20	5083
22	52	12	99	24	34	21	0	11	5048
13	36	10	94	15	23	16	4	11	5046
29	27	3	97	9	30	15	0	22	5061
17	48	14	120	18	34	13	6	20	5036
14	35	9	105	14	23	13	4	23	5029
22	46	26	88	26	42	24	3	24	5017
2	22	4	48	8	11	4	5	7	1430
22	9	18	10	18	17	19	7	17	5060
11	19	15	10	15	18	16	12	23	5031
12	19	17	8	16	13	19	9	11	4974
4	13	18	18	10	9	18	4	21	5038
9	20	13	9	19	16	16	10	10	5040
12	17	13	30	6	16	19	7	14	5038
13	21	19	31	14	17	15	16	20	5048
19	15	21	30	9	9	26	5	13	5018
9	18	9	16	11	12	15	12	13	5023
14	14	11	7	9	13	17	10	21	5033
4	27	14	11	6	6	25	6	8	5031
14	21	8	44	15	8	13	11	9	5000
10	17	16	14	12	9	18	17	10	4951
10	17	22	7	11	4	12	12	5	4997
8	18	22	17	10	10	9	7	8	4954
14	17	12	11	11	1	18	16	12	4963

Rank			1	2	3	4	5
Author	Texts	Segment	the	of	to	a	and
Joe Klein	Public Lives	17	288	129	140	126	118
Joe Klein	Public Lives	18	310	108	135	158	102
Joe Klein	Public Lives	19	314	136	143	130	105
Joe Klein	Public Lives	20	275	97	125	132	114
Joe Klein	Public Lives	21	280	127	152	136	116
Joe Klein	Public Lives	22	326	139	120	153	97
Joe Klein	Public Lives	23	296	127	138	139	108
Joe Klein	Public Lives	24	291	151	137	131	111
Joe Klein	Public Lives	25	311	154	130	112	102
Joe Klein	Public Lives	26	335	158	127	132	120
Joe Klein	Public Lives	27	316	149	119	134	120
Joe Klein	Public Lives	28	302	138	142	122	107
Joe Klein	Public Lives	29	288	128	113	129	79
Michael Kelly	New Republic	1	353	151	171	119	140
Michael Kelly	New Republic	2	381	185	170	123	120
Michael Kelly	New Republic	3	350	145	138	124	108
Michael Kelly	New Republic	4	354	166	155	106	140
Michael Kelly	New Republic	5	313	161	147	117	147
Michael Kelly	New Republic	6	360	136	126	123	131
Michael Kelly	New Republic	7	394	178	164	110	119
Michael Kelly	New Republic	8	353	183	143	123	148
Michael Kelly	New Republic	9	331	157	118	129	135
Michael Kelly	New Republic	10	355	187	106	158	158
Michael Kelly	New Republic	11	402	151	124	117	166
Michael Kelly	New Republic	12	365	179	85	118	185
Michael Kelly	New Republic	13	367	181	99	149	162
Michael Kelly	New Republic	14	383	173	102	140	150
Michael Kelly	New York Times	15	257	151	122	154	150
Michael Kelly	New York Times	16	338	145	128	117	134
Michael Kelly	New York Times	17	372	198	132	143	139
Michael Kelly	New York Times	18	370	179	152	127	151
Michael Kelly	New York Times	19	330	155	143	122	149
Michael Kelly	New York Times	20	345	127	152	129	117
Michael Kelly	New York Times	21	339	154	119	130	152
Michael Kelly	New York Times	22	335	128	136	130	113
Michael Kelly	New York Times	23	304	161	135	148	101
Michael Kelly	New York Times	24	342	163	153	148	115
Michael Kelly	New York Times	25	309	144	166	126	121
Michael Kelly	New York Times	26	309	189	127	158	133
Michael Kelly	New York Times	27	327	154	153	158	128
Michael Kelly	New York Times	28	81	54	34	52	49

6	7	8	9	10	11	12	13	14
in	is	that	he	it	not	for	was	on
72	93	58	35	63	67	45	28	22
82	79	59	64	48	56	52	25	32
81	69	61	27	52	50	44	31	36
86	120	44	49	64	66	42	36	28
81	80	51	37	49	57	39	32	34
110	72	41	57	41	49	35	50	32
79	72	50	45	54	45	45	38	19
95	81	57	41	57	36	44	35	25
68	80	69	26	67	59	57	37	41
80	79	53	61	51	52	39	48	25
75	62	45	39	58	49	41	37	27
85	73	65	91	55	54	46	30	41
69	63	40	36	43	33	52	53	35
98	80	74	65	46	45	52	39	24
191	59	101	32	51	44	38	57	30
99	57	94	67	36	42	57	48	38
102	68	85	42	36	52	40	31	40
97	56	98	42	57	41	38	42	32
100	99	100	40	59	60	52	41	40
66	55	80	48	70	48	32	45	34
96	74	76	25	41	42	36	27	27
94	120	86	72	63	40	48	47	40
108	37	54	50	44	35	28	37	34
111	60	34	30	41	43	39	37	29
130	32	36	40	42	27	23	65	45
106	50	54	31	30	31	30	42	50
97	61	32	36	32	40	41	41	40
112	65	74	65	32	42	42	60	20
108	11	82	79	34	19	48	79	37
104	91	74	76	44	37	40	32	41
94	41	88	60	55	25	35	86	39
106	52	78	91	58	35	29	53	44
96	45	99	60	33	36	40	46	46
88	63	113	40	26	26	62	21	45
72	44	86	60	28	35	46	37	49
84	44	73	18	44	34	62	42	51
96	56	68	55	35	42	32	40	55
98	48	72	60	23	23	58	38	45
99	79	49	65	37	28	49	37	39
110	46	64	79	31	17	42	33	34
34	21	16	24	8	10	17	19	7

15	16	17	18	19	20	21	22	23
his	as	but	with	be	are	by	this	has
24	21	43	25	46	47	26	18	26
32	29	41	32	48	29	23	27	35
24	21	42	30	42	25	25	30	29
22	28	38	27	41	27	29	33	29
36	29	36	26	47	42	23	20	21
26	49	32	34	26	41	25	24	38
23	20	36	19	40	27	17	22	31
31	41	47	20	29	31	30	20	41
21	28	41	27	41	39	21	20	25
41	33	29	19	28	20	29	23	35
36	33	26	30	48	33	36	15	21
42	29	43	22	15	21	14	28	46
36	21	28	29	22	17	17	29	21
21	28	19	30	25	18	45	32	20
20	43	24	19	24	21	36	29	20
54	29	27	28	21	15	21	30	11
24	30	30	31	28	23	40	24	20
36	32	17	13	22	8	39	28	21
22	46	34	21	24	29	22	34	19
35	33	23	25	28	16	28	31	18
20	21	27	36	21	26	27	40	23
30	51	23	22	26	13	29	33	21
33	23	12	36	8	34	13	21	12
20	20	19	34	16	38	24	32	15
34	29	20	55	9	6	15	18	2
26	25	26	35	12	15	27	21	11
24	28	29	57	22	26	21	21	16
53	34	25	38	19	12	18	20	22
78	36	21	32	21	2	34	13	6
45	41	32	24	16	24	28	19	12
32	33	22	28	18	13	31	29	11
32	51	25	35	18	17	29	29	19
42	52	38	38	28	19	25	20	18
26	38	27	35	21	21	29	16	28
27	43	22	44	25	18	40	13	8
12	43	31	43	26	19	31	12	27
35	37	19	37	10	20	20	22	16
36	55	24	44	22	18	20	14	24
37	42	21	39	29	18	22	19	20
49	37	19	39	13	11	24	15	23
5	17	6	14	3	9	3	12	7

24	25	26	27	28	29	30	31	32
have	I	they	who	had	at	an	would	will
27	14	42	24	11	18	22	18	19
20	7	31	16	10	18	28	17	44
33	8	25	21	11	11	23	28	15
20	15	20	12	11	8	10	22	24
48	21	25	18	15	14	20	17	24
27	15	21	22	15	13	23	10	20
39	17	35	14	20	20	18	22	22
27	8	14	28	10	16	19	11	10
38	17	27	13	9	15	18	17	19
22	18	17	15	9	11	17	16	29
21	10	11	15	16	24	16	21	20
12	17	21	20	18	14	17	18	26
24	7	11	28	17	13	18	17	8
32	19	19	19	11	12	21	28	18
9	8	18	26	20	15	16	16	5
12	25	10	18	19	33	16	13	21
29	19	21	9	11	15	15	28	10
12	19	12	14	26	23	14	20	15
15	4	22	22	29	13	14	13	12
21	17	10	22	17	26	15	11	30
23	39	32	21	17	31	16	10	8
28	32	23	25	27	26	21	25	9
15	33	28	24	21	27	15	9	4
21	18	37	19	31	21	8	12	3
8	28	48	9	60	30	12	19	2
7	25	27	10	45	33	18	10	9
12	25	30	11	24	23	15	6	24
20	24	7	21	32	20	20	17	5
7	11	4	19	38	27	12	24	2
19	29	34	18	35	14	22	17	7
15	25	11	23	38	16	20	14	3
17	49	7	15	25	25	17	15	1
18	26	22	17	23	12	14	20	25
20	29	18	27	26	25	21	11	8
21	36	30	23	27	14	20	21	16
16	23	16	29	27	26	32	19	12
9	39	11	28	30	24	19	15	13
20	24	22	28	24	27	19	23	13
11	31	17	26	18	18	18	8	25
22	21	7	23	15	35	22	11	14
6	14	5	13	3	9	8	2	6

33	34	35	36	37	38	39	40	41
there	from	said	president	about	more	one	we	or
23	19	9	5	22	15	14	13	24
12	9	2	16	13	31	19	10	12
22	11	5	9	13	31	20	24	21
22	18	10	15	19	23	14	16	16
37	9	11	13	13	22	15	12	12
17	17	9	6	20	27	9	13	14
16	16	8	7	13	30	13	21	18
15	22	5	10	19	20	24	18	16
23	15	5	9	21	19	15	14	11
15	19	10	13	14	27	10	11	16
24	19	8	15	10	28	17	8	12
14	18	19	12	32	16	16	19	16
17	19	10	14	18	18	14	11	12
15	17	11	18	1	7	8	13	15
9	17	12	16	6	8	11	9	18
11	20	11	31	9	10	12	13	14
10	11	7	22	9	12	9	11	20
11	19	5	23	23	6	7	9	23
10	19	14	28	11	14	21	15	27
11	14	5	32	28	9	15	10	29
14	21	6	17	13	11	12	9	18
22	14	14	13	19	13	13	6	14
10	33	24	2	14	21	20	21	26
20	22	29	0	10	4	20	20	17
18	26	32	0	6	11	34	14	21
17	25	21	0	7	2	36	41	7
14	20	20	1	9	16	19	26	20
9	12	10	20	20	12	17	12	16
3	17	5	3	8	10	15	8	8
10	18	9	30	7	12	15	1	12
14	15	3	17	13	10	15	16	11
15	11	4	20	20	10	7	16	7
9	13	42	43	13	11	18	14	13
17	16	45	24	12	12	10	25	16
16	25	63	30	12	11	17	23	15
13	21	39	15	16	7	14	12	18
7	15	50	25	15	7	10	10	4
13	16	43	35	13	12	13	10	10
11	14	32	28	9	16	13	4	11
4	24	29	27	12	11	14	9	12
4	7	16	6	2	6	1	7	3

42	43	44	45	46	47	48	49	50	Segment
were	do	been	you	all	what	if	their	so	Length
23	18	17	15	8	9	21	25	11	5054
4	8	12	8	7	5	16	12	10	5098
18	12	10	7	13	9	16	19	7	5107
5	27	10	7	16	12	16	24	7	5066
16	15	15	21	8	11	16	11	13	5074
15	16	17	9	10	8	13	11	15	5070
6	7	14	15	7	11	20	14	14	5067
18	17	13	10	9	13	10	8	5	5127
11	16	17	14	10	9	13	17	9	5157
12	12	11	9	12	8	25	16	13	5125
14	8	9	8	13	11	22	11	6	5175
17	11	16	7	22	21	14	10	11	5131
11	5	11	6	11	3	8	15	10	4617
11	10	9	9	9	15	15	15	10	5055
12	8	14	4	10	21	8	14	6	5063
10	7	5	23	15	11	6	8	12	5064
16	18	13	8	26	10	5	9	18	5080
15	15	9	10	19	17	10	4	8	5060
24	14	9	13	18	25	6	15	7	5058
17	9	8	11	14	9	10	7	9	5064
19	11	10	8	22	24	6	9	19	5067
8	9	16	13	15	14	13	9	11	5039
22	10	5	21	14	12	11	15	11	5076
31	8	14	7	23	7	4	14	10	5064
40	1	26	20	11	11	1	13	9	5080
18	4	19	8	24	7	5	26	12	5061
26	14	6	14	21	11	8	18	12	5060
18	13	8	4	11	13	6	7	16	5033
18	5	8	5	14	9	6	8	9	5067
5	10	15	14	12	26	11	10	13	5041
18	13	12	17	11	28	6	7	6	4990
11	15	13	18	12	21	2	7	15	4983
11	20	10	16	10	17	10	5	9	5045
13	9	18	5	10	15	7	11	3	5044
22	27	14	16	4	16	11	7	3	5060
10	5	17	10	10	9	7	14	4	5050
8	12	8	10	11	16	3	9	10	5061
14	9	13	8	4	8	6	11	9	5044
7	12	10	18	9	16	7	8	4	5084
6	4	5	9	9	10	10	11	8	5049