QUANTITATIVE AUTHORSHIP ATTRIBUTION: A HISTORY AND AN EVALUATION OF TECHNIQUES

JACK WILLIAM GRIEVE

BACHELOR OF ARTS, SIMON FRASER UNIVERSITY, 2002

A thesis submitted in partial fulfillment
of the requirements for the degree of

MASTER OF ARTS

IN THE DEPARTMENT
OF
LINGUISTICS

SIMON FRASER UNIVERSITY
Summer 2005

APPROVAL

Name: **Jack William Grieve**

Degree: **Master of Arts**

Title of Thesis: **Quantitative Authorship Attribution:
A History and an Evaluation of Techniques**

Examining Committee:

Dr. Zita McRobbie
Chair
Associate Professor, Department of Linguistics

Dr. Paul McFetridge
Senior Supervisor
Associate Professor, Department of Linguistics

Dr. María Teresa Taboada
Supervisor
Assistant Professor, Department of Linguistics

Dr. Fred Popowich
External Examiner
Professor, School of Computing Science

Date Defended: June 17, 2005

ABSTRACT

I here present a history of the field of quantitative authorship attribution and an evaluation of its techniques. The basic assumption of quantitative authorship attribution is that the author of a text can be selected from a set of possible authors by comparing the values of textual measurements in that text to their corresponding values in each author's writing sample. Over the centuries, many measurements have been proposed, but never before have the majority of these measurements been tested on the same dataset. Until now investigators of authorship have not known which measurements are the best indicators of authorship. Such information is crucial if our current techniques are to be used effectively and if new more powerful techniques are to be developed. Based on the results of this study, I propose that the best approach to quantitative authorship attribution involves the analysis of many different types of textual measurements.

ACKNOWLEDGEMENTS

I would like to thank Paul McFetridge, Maite Taboada, E. W. Roberts, Joseph Rudman, Graeme Trousdale, Ivan Zubov, Fred Popowich and most especially Tom Grieve and Paula Chmilar for their help and support.

CONTENTS

TABLES

FIGURES

1 INTRODUCTION

Everyone knows that language is variable. Two individuals of the same generation and locality, speaking precisely the same dialect and moving in the same social circles, are never absolutely at one in their speech habits.

Edward Sapir (1949:147)

When we read a text we consider the source. We trust, respect and obey the words of some authors more than others, even if their words are exactly the same. Knowledge of source allows us to judge the information that a text contains, in light of what we know about the personality, reputation and wisdom of its author. But sometimes the source of a text is unknown. This is the problem of authorship attribution: how can one discover the author of an anonymous text?

The universe of valid authorial evidence consists of internal evidence and external evidence. Internal evidence includes all those clues of authorship found within the string of characters—e.g. graphemes (*A, B, ..., Z*), digits (*0, 1, ..., 9*), whitespace (space, paragraph, tab), symbols (*$, @, &, ...*) and punctuation marks (*. , : ; ...*)—that make up a text. External evidence includes everything else. Usually, there is a sufficient amount of external evidence to allow a reader to attribute a text: the author may have handed the reader the document, or the reader may recognize the handwriting in which the text was written, or there may be only one other person who has access to the place where the text was found. But if an analysis of external evidence does not reveal the author of a text, then the text can only be attributed through an analysis of its own internal evidence. This is all the evidence that is left.

A direct analysis of a text will often lead to a quick attribution: the text may contain a reference to the author, or an opinion that could only have come from one possible source. But such direct internal evidence may also be absent or unreliable or intentionally misleading. In such cases, the investigator may compare the anonymous text to a series of possible author writing samples, in order to determine which writing sample is the best match. In qualitative authorship attribution, a person, preferably an expert, reads through the anonymous text and each possible author's writing sample in search of distinctive patterns of style and theme. A successful qualitative authorship attribution depends on the investigator's ability to read texts and uncover evidence of authorship. In quantitative authorship attribution, a person, or preferably a computer, compares the values of certain textual measurements in the anonymous text to their values in each possible author's writing sample. A successful quantitative authorship attribution depends on the investigator's selection of a set of textual measurements whose values are relatively consistent across each possible author's writing sample and relatively variable across the set of possible authors. Both approaches can lead to successful attributions, but the advantages of quantitative authorship attribution are clear: only quantitative techniques may be empirically evaluated and mechanically applied.

The goal of this thesis is to identify and compare the value of the most commonly used sets of textual measurements in quantitative authorship attribution. Until now, there has never been a large-scale comparison of the many quantitative indicators of authorship that have been proposed over the past two centuries. Such a comparison is long overdue: if investigators are to resolve current cases of disputed authorship and develop new and more powerful techniques, then investigators must be aware of which of our current measurements are the best indicators of authorship. This project fulfills the need for a comprehensive comparison of textual measurements, but also makes several other important contributions to the field of quantitative authorship attribution. First, the literature review presented here is the most extensive and accurate history of the field that has ever been compiled. Second, I discuss in detail how to evaluate the potential of a textual measurement—a topic that has been ignored far too often in the past, but which must be addressed here if accurate results are to be obtained. Third, in addition to testing current techniques, I also introduce and test new textual measurements. Finally, I demonstrate that the best approach to quantitative authorship attribution is one that takes into consideration the values of as many different types of textual measurements as possible—a conclusion that challenges most quantitative authorship attribution studies, which only analyze the values of one type of textual measurement.

This thesis is organized as follows. In Chapter 2, I present a history of the field of quantitative authorship attribution. From this history, a set of basic textual measurements is identified. In Chapter 3, I explain how the values of each of these textual measurements is calculated for any text. I also describe how the texts are prepared for analysis, how the values of the textual measurements are compared, and how the results of these comparisons are used to attribute a text. Based on this information, anyone with minimal programming knowledge, or enough time to count through texts by hand, should be able to implement any of the methods tested in this study. In Chapter 4, I describe the study's experimental design: I explain how the dataset—the corpus of possible authors—was compiled and how it was used to gauge the performance of each set of textual measurements. In Chapter 5, I present the results of these tests and discuss their significance to the field of quantitative authorship attribution. Based on these results, I propose a general procedure for quantitative authorship attribution that involves analyzing the values of many types of textual measurements.

2 HISTORY

2.1 INTRODUCTION[1]

For centuries there has been interest in developing quantitative methods for determining the author of anonymous texts. The basic assumption of quantitative authorship attribution is that the author of a text can be chosen from a set of possible authors by comparing the values of a set of textual measurements in that text to their corresponding values in each possible author's writing sample. A successful quantitative authorship attribution therefore depends on the investigator's selection of a set of textual measurements that is capable of distinguishing between that particular set of possible authors. The history of quantitative authorship attribution is the history of the search for these characteristic measurements of authorship.

2.2 METER & RHYME[2]

The earliest quantitative studies of authorship involved the analysis of meter and rhyme. One of the first of these studies was published in 1787 by Edmond Malone, an expert on the chronology and provenance of Shakespeare's plays. who argued that Shakespeare did not write any of the three parts of *Henry VI,* based on its author's frequent use of end-stopped lines and infrequent use of double endings and rhyming lines—metrical properties that are not characteristic of Shakespeare's plays. A second early quantitative study of authorship was published in 1812, when Henry Weber attributed parts of *The*

[1] The main sources for this history of quantitative authorship attribution are Holmes (1985, 1994, 1998), McMenamin (1993, 2002), Williams (1970), Erdman & Fogel (1966), Bailey (1969), Kenny (1982), Love (2002), Oakes (1998), Oakman (1980), and McEnery & Oakes (2000).

[2] For general metrical references see Erdman & Fogel (1966), Williams (1970), and Chambers (1930).

Two Noble Kinsmen to William Shakespeare and parts to John Fletcher based on the number of feminine endings in different parts of the play. Then, in an influential 1850 study, James Spedding made a similar case for Fletcher as coauthor of Shakespeare's *Henry VIII* based on the frequency of feminine endings, which, as Richard Roderick had first observed in 1758, occur more frequently in *Henry VIII* than in any other of Shakespeare's plays.

Impressed by these early attempts to quantify authorship, most especially Spedding's seminal paper, the British philologist F. J. Furnivall—one of the founders and original editors of the Oxford English Dictionary—established the New Shakespeare Society in 1873. One of the society's main goals was to promote the use of quantitative "stylometric" techniques to help resolve the many cases of disputed authorship and chronology that plagued the Shakespearean Canon. A number of quantitative authorship studies appeared in the *Transactions of the New Shakespeare Society*. For example, John K. Ingram argued in the pages of the journal's first edition (1874) that Spedding's conclusion was confirmed by an analysis of the variance in the frequency of light and weak endings in *Henry VIII.* Later, in the 1886 edition of the *Transactions,* Robert Boyle pushed Spedding's theory even further by using metrical and thematic evidence to argue that Shakespeare had no part in *Henry VIII* at all—a conclusion that we now believe to be false, though it was endorsed then by Robert Browning. But the society's most outspoken and controversial member was undoubtedly the Reverend F. G. Fleay, whose absolute confidence in the use of statistics and metrical evidence, as expressed in the society's first published paper (1874a), even made Furnivall and the more conservative members of the Society uneasy. Indeed, they were probably right: in the same edition of the *Transactions* (1874b), Fleay used his metrical tests to show that Shakespeare had little to do with the writing of *The Taming of the Shrew*—another conclusion that we now believe to be false.

A particularly good example of the type of quantitative evidence that Fleay and other early investigators of authorship used to distinguish the work of Elizabethan playwrights can be found in Fleay's *Shakespeare Manual* (1876:69-72). Here Fleay provided a number of metrical tests that he claimed were capable of distinguishing between the works of Shakespeare and his contemporaries. For example:

> Shakespeare admits, in addition to the regular 5-foot blank verse line, the Alexandrine, short lines of 1, 2, or 3 feet, and rhyming lines of 4 or 5 feet.
> Fletcher can be at once distinguished by the number of female lines, in which he exceeds every other English author.
> Massinger is known instantly by his numerous weak endings, in which he indulges beyond any other writer.

In the *Shakespeare Manual,* Fleay also provided tables containing the raw Shakespearean data upon which his metrical tests were based: counts of the number of lines; the proportion of lines of blank verse, rhymed verse, and prose; the number of unusually long and short lines; and the number of redundant syllables. But, as the Shakespearean scholar Sir Edmund Chambers remarks (1930:256), however useful such information may be, the tables are often incorrect. Because of their questionable data and conclusions, the quantitative approaches to authorship attribution introduced by these early investigators were rejected by most literary scholars of the time. Perhaps the most scathing attack came from the poet Algernon Swinburne who penned "Report of the Proceedings on the First Anniversary Session of the Newest Shakespeare Society" (1876)—a parody of the Society and their use of such methods as the always-reliable "heavy-monosyllabic-eleventh-double ending test." However, it should be noted that Swinburne was no impartial observer: how could he not have taken offence when Fleay asked, in the Society's original publication (1874a:2), "is not the trick of Swinburne's melody easily acquired and reproduced?"

While interest in quantitative authorship attribution has grown since the late nineteenth century, metrical tests are far less popular today than they once were. For a large part, this is probably because most modern investigators of authorship have focused on prose rather than on poetry or drama, where rhythm is perceived to be a more suitable indicator of authorship. But metrical indicators have continued to be used in Shakespearean studies, most notably in Don Foster's (1989) attribution of an obscure poem entitled *A Funeral Elegy,* by W. S.. The *Elegy* was published in 1612 in remembrance of William Peter—an Englishman who had been murdered earlier that year by John and Edward Drew after having spent an afternoon with the brothers carousing through the pubs of Exeter. Intrigued by the *Elegy*'s signature and its similarity in substance and style to the works of William Shakespeare, Foster compared the *Elegy* to the writings of the Bard and William Strachey—the only other Jacobean writer who

Foster believed could have possibly written the anonymous text. The attribution was based on a variety of quantitative and qualitative tests, including such stylometric measurements as the frequency of unstopped, feminine and open lines. At that time, Foster concluded that Shakespeare was the *Elegy*'s most likely author.

Recent metrical authorship studies of the Shakespearean Canon have also been conducted by the Claremont Shakespeare Clinic—a group of students led by the political scientist Ward Elliott and the mathematician Robert J. Valenza. As Don Foster wrote (1996:247):

> Ward Elliott founded the Shakespeare Clinic in 1987, having received his call from an association of anti-Stratfordians called the Shakespeare Authorship Round Table. His original "assignment" was to find Shakespeare's ghost, winnowing likely candidates one by one until the true "Shakespeare" emerged.

To this end, the members of the Clinic compiled a large corpus of Shakespeare's plays and poems, works of disputed Shakespearean authorship (e.g. *A Funeral Elegy*), and texts written by the main Shakespeare claimants (e.g. Francis Bacon, Christopher Marlowe, Sir Walter Raleigh, Queen Elizabeth). Elliott and Valenza then used a battery of tests to compare the various claimants to Shakespeare. These tests included both traditional metrical tests, such as the percentage of feminine endings and open lines, and novel metrical tests, such as the metrical filler test, which is based on the frequency of meter preserving words like *the* and *that* in the phrases "to *the* which place" and "if *that* the world was young." Elliott and Valenza also used the leaning microphrase test, proposed by Tarlinskaja (1987), that is based on the frequency of lines with "clinging monosyllables," which are unstressed because of their metrical position in the line. The Clinic published their final results in a lengthy 1996 article "And Then There Were None: Winnowing The Shakespeare Claimants" in the journal *Computers and the Humanities* (*CHum*). They concluded, with palpable disappointment, that the only true Shakespeare was Shakespeare himself: all of the claimants failed their tests. But Elliott and Valenza's research did lead to one controversial conclusion: their tests found that Shakespeare did not write the *Funeral Elegy*.

It did not take long for Foster to respond: the editors of *CHum* allowed Foster to publish a "response" to the Elliott and Valenza article, which appeared in the same edition of the journal. Foster did not concern himself directly with the Clinic's analysis of the *Funeral Elegy*. Instead he presented a damning critique of their entire program, focusing in particular on the validity of their tests. For example, Foster argued that the leaning microphrase test was subjective and far more sensitive to changes in prosody and register than to changes in authorship. Foster also took time in his critique to explain his relationship to the Claremont Shakespeare Clinic (1996: 254):

> Little of what I have said here will come as news to Ward Elliott. I served as an advisor to the Shakespeare Clinic for several years, did my best to help, but finally withdrew, exasperated that problems with accuracy and with the validity of testing were never addressed, only multiplied.

Elliott and Valenza responded to this attack in far too many publications to review here, but their official response "The Professor Doth Protest Too Much, Methinks" came in a 1999 edition of *CHum*. In this article they defended their tests and accused Foster of attacking their work only because they had called his attribution of the *Funeral Elegy* into question. Elliott and Valenza recounted how the Claremont Clinic had been in communication with Foster's "one-man show" at Vassar, until the Clinic's tests started to disprove Foster's famous attribution, at which point Foster resigned, promising that the Clinic's reputation would be ruined if they ever published their results. The debate did not stop here: the editors of *CHum* allowed Foster (1999) to publish another back to back response, which further enraged Elliott and Valenza who fired back with counterattacks including a 2002 reply, "So Many Hardballs So Few Over the Plate," and a 2001 paper in which they claim to prove that Shakespeare did not write the *Funeral Elegy*.

The final chapter of this story came in 2002 when Elliott and Valenza's attacks were validated: Don Foster withdrew his attribution of the *Funeral Elegy* (see Niederkorn 2002), and accepted the attribution of the *Elegy* to John Ford, as proposed by Gilles Monsarrat (2002) and Brian Vickers (2002).[3]

2.3 WORD-LENGTH

In 1851, a year after Spedding's seminal article was published, Augustus de Morgan—the famous mathematician immortalized by the laws of set theory that bear his name—proposed a new indicator of authorship. In a letter written to the Reverend W. Heald, de

[3] For other more metrical studies, see Merriam (2000), who analyzed the authorship of *Edward III* using Burrows' method and cusum charts to analyze metrical textual variables (lines with feminine endings, lines with more or less than ten syllables, lines with more or less than five stresses).

Morgan suggested that one could discover if the Epistles of Saint Paul had truly been written by a single man by comparing their average word-lengths. While it appears that neither the Reverend nor the Professor ever tested the theory out, the letter was published posthumously, thirty years later, in a memoir compiled by Sophia de Morgan in honor of her late husband (1882:215-216).[4] Soon after it was published, T. C. Mendenhall—an American geophysicist at Ohio State University—read the letter and began to conduct attribution experiments. Mendenhall expanded de Morgan's original technique so as to examine a text's entire word-length frequency distributions—the number of one-grapheme words, two-grapheme words, three-grapheme words, etc. in a text. The distributions he found reminded him of the unique spectrum of light that each of the chemical elements emits, and so he dubbed an author's word-length frequency distribution a *word-spectrum.* Mendenhall's research is described in his two classic papers: "The Characteristic Curves of Composition" (1887) and "A Mechanical Solution to a Literary Problem" (1901).

In the 1887 study, Mendenhall conducted three separate experiments. First, he compared the word-spectra of samples of Dickens' *Oliver Twist* and Thackeray's *Vanity Fair.* Mendenhall found that, while the spectra appeared to be stable for each of these authors, they were also disappointingly similar.[5] As Mendenhall wrote (1887:241):

> Although these curves differ, and while it is believed that the differences will persist with an increased number of words, it is certainly surprising, that in the analysis of ten thousand words from Dickens, and the same from Thackeray, so close an agreement should be found.

In the second experiment, Mendenhall compared two John Stuart Mill essays, "Political Economy" and "Essay on Liberty", in order to test "the persistence of form in compositions belonging to different periods of the author's life, and upon different subjects" (1887:242). While Mendenhall was pleased to find that both of Mill's spectra did share the same unusual two-letter word-length peak, overall, he had to admit that the two spectra had less in common than the spectra found in the samples of Dickens' and Thackeray's works. Finally, Mendenhall compared two speeches by Edward Atkinson—

[4] The letter is reprinted in Lord (1958) and Williams (1970).

[5] For example, both authors' word-spectra peak at 3-letters (i.e. both authors favor three letter words) and the top six ranked word-lengths for each of the authors is identical (three-, four-, two-, five-, six-, and seven-letter words).

"Address to Workingmen" and "To Alumni of Theological Seminary"—to see if two texts which share the same subject and the same author, but which were delivered to different audiences, would differ. The results were as one might expect: the more formal lecture contained more words over six letters.

The results of all three of these experiments suggest that word-length is not a good indicator of authorship. Indeed, at the end of the 1887 paper, Mendenhall clearly states that word-length distributions would only be a useful measurement in some cases of authorship elimination:

> If the two compositions should produce curves which are practically identical, the proof of a common origin would be less convincing; for it is possible, although not probable, that two writers might show identical characteristic curves.

Regardless, Mendenhall continued to experiment with word-length as an indicator of authorship. Bankrolled now by a wealthy Boston Philanthropist named Augustus Hemingway, Mendenhall set out, along with a pair of secretaries and a primitive adding machine, to determine whether or not Francis Bacon was the true William Shakespeare. After analyzing 400,000 words of Shakespeare and 200,000 words of Bacon, Mendenhall and his team reported the results of their research (1901): Shakespeare and Bacon were characterized by very different word-length distributions. The most obvious difference was that whereas Bacon's spectrum peaked, like most authors, at three letters, Shakespeare's spectrum peaked at four. Mendenhall could find this peak in the texts of only one other man—Christopher Marlowe. As Mendenhall wrote: "Christopher Marlowe agrees with Shakespeare about as well as Shakespeare agrees with himself" (1901:105).

Mendenhall's method has not been adopted by many modern investigators of authorship. This is because, as experts such as C. B. Williams (1970), M. W. A. Smith (1983) and, indeed, Mendenhall himself (1887) have demonstrated, the measurement is far more sensitive to differences in register, subject and language than to differences in author. But there have been a few notable attempts to use word-length as an indicator of authorship. First, in a 1963 study, C.S. Brinegar examined the word-length distributions of the ten Quintus Curtius Snodgrass Letters—a series of letters which were published in 1861 editions of New Orleans' *The Daily Crescent* and which were thought to have been written by Mark Twain. Two major pieces of external evidence point to Twain: he was known to have been working in New Orleans as a Mississippi steamboat pilot at the time, and he was known to have used the pen name "Thomas Jefferson Snodgrass" in some of his contributions to the Keokuk *Post* of Iowa. The letters are of historical significance because if they are real they would shed some light on Twain's uncertain role in the American Civil War: it would seem that he had served briefly as a Confederate volunteer when the Mississippi River was closed at the outset of the war, but he was also accused of being a deserter, for it is known that he had left for the Nevada silver mines later that year with his brother, who had been appointed Secretary of the newly formed territory by no less a Unionist than President Abraham Lincoln himself. To resolve this case of disputed authorship, Brinegar used Mendenhall's method to compare the Snodgrass letters to other letters known to have been written by Twain around the same time. The only adjustment Brinegar made to the method was that he used the two-sample t-test and the Chi-square test to compare the word-spectra of these texts, as opposed to a simple visual comparison.[6] Based on significant differences between the word spectra of the texts, Brinegar concluded that Twain had not written the Snodgrass letters. But it is important to note that even if Twain and Snodgrass were found to exhibit similar curves, Brinegar would not have been justified in concluding that the letters were written by the same author, as there is no reason to assume that another possible author did not produce similar patterns in their texts as well.

A second series of studies is also worth mentioning because of the unique way in which word-length was measured. In a number of papers authored or co-authored by the German physicist Wilhelm Fucks (Fucks 1952, Fucks 1954, Fucks & Lauter 1965), the potential of a syllable-based word-length distribution for authorship attribution was explored. In the 1952 and 1954 papers, Fucks evaluated three syllable-based indicators: average word-length in syllables, word-length frequency distribution in syllables (i.e. proportion of word-tokens with one syllable, proportion with two syllables, etc.), and the average distance between *i*-syllable words (i.e. mean distance between two one syllable word-tokens, two two syllable word-tokens, etc.). At first these techniques appeared to have some potential, but as Holmes (1985:331) reports, based on an analysis of German

[6] The two-sample t-test is a simple statistical test used to determine if two population means are equal, and the Chi-square test is a simple non-parametric goodness of fit test used to compare a set of observed and a set of expected frequencies.

and English authors, Fucks & Lauter (1965) concluded that word-length measurements were better indicators of languages than of authors.

Finally, Forsyth, Holmes and Tse (1999) used a more complex syllable-based word-length measurement in their investigation of the authorship of the *Consolatio Ciceronis*—a philosophical treatise written by the great Roman orator Cicero in 45 BC upon the death of his daughter. The problem that this text poses is that while no complete copies of the *Consolatio* were known to have survived the fall of Rome, a text was published in 1583 whose editor, Carlo Sigonio, claimed was a complete copy of Cicero's original text, though he was not always believed. To investigate this case of disputed authorship, Forsyth *et al.* measured the percentage of one-, two-, three-, four-, five- and six-syllable words in texts written by six classical Latin authors (including Cicero), and by five Neo-Latin authors (including Sigonio), but also measured the percentage of one-, two-, three- and four-syllable words that are followed by one-, two-, three- and four-syllable words. They then used these syllable-based measurements to see if either Cicero or Sigonio had written the sixteenth century edition of the *Consolatio*. Forsyth *et al.* found that the additions to the text were quite unlike Cicero's known works and far more similar to Neo-Latin writers, most especially Sigonio.[7]

2.4 Sentence-Length

Following the publication of Mendenhall's 1887 study in *Science*, the journal published a series of letters commenting on his research. In the first letter, H. T. Eddy (1887) praised Mendenhall for his novel technique, but offered some alternative textual measurements that he believed might be even better indicators of authorship, including a text's average sentence-length and sentence-length distribution (e.g. the number of sentences in a text which contain one to five words, six to ten words, etc.). The next three letters were published in rapid succession three years later. The first of these letters, written by A.B.M., described the results of an experiment of the type Eddy proposed, and included a hand-drawn graph charting the distinctive sentence-length distributions of works by

[7] For other word-length studies, see Radday (1970) for an authorship study of the Book of Isaiah that considers both syllable- and grapheme-based word-length measurements, in addition to other evidence (see sections 2.4, 2.14); and see Foster (1989), where average word-length is one of the many indicators used to attribute the *Funeral Elegy* (see section 2.2).

Thomas Carlyle, Thomas De Quincey, and Samuel Johnson. A.B.M.'s letter may constitute a case of disputed authorship in and of itself: while the initials do not quite match up, Richard W. Bailey suggested that T. C. Mendenhall "may, indeed, have written the initialed letter" (1969:218)—in order, one assumes, to rekindle interest in his own work.[8] If this was the case, Mendenhall succeeded, for the journal soon published a third letter, this time by H. A. Parker, who verified A.B.M.'s sentence-length distribution for Carlyle's *French Revolution*, but showed that it did not agree with the curve of Carlyle's *Sartor Resartus*. Parker concluded: "this goes to show, if it does not prove, that for detective purposes the method is valueless." In the fourth letter, presumably written by A.B.M., though it is signed with only an M., the author explains curtly to Parker that "a comparison of three hundred sentences proves nothing one way or another."

In 1888, William Benjamin Smith—head of the Tulane math department, and noted poet, translator, Christian philosopher, social Darwinist and white supremacist—published, under the pseudonym Conrad Mascol, a two-part quantitative authorship attribution study of the Pauline Epistles that considered the average sentence-length of the texts.[9] Instead of measuring average sentence-length in words, like Eddy had suggested, Smith measured the average number of sentences per page. This is a more sensitive measurement: if a page of text is defined as a fixed number of characters, then two texts with the same average sentence length may not have the same average number of sentences per page, depending on their average word-lengths. In essence, Smith was measuring a text's average sentence-length in characters. Smith used this method, as well as others (see sections 2.5 and 2.11), to test the consistency of seven of the Pauline Epistles: he compared the four core Epistles that most modern theologians attribute to Paul (*Romans, I and II Corinthians,* and *Galatians*) to three Epistles whose true authorship has been in question for centuries (*Ephesians, Philippians, Colossians*). Smith concluded that the two groups of texts were most likely written by two different Greek authors.[10]

[8] Bailey provides no specific evidence, but the letter's graph is similar in style to those found in Mendenhall's original paper, but the initials do not match, and the letter bears a London and not a Nebraskan dateline.

[9] Smith's work was reintroduced to modern authorship attribution by Rudman (1998).

[10] See also the pioneering work of Lucius Sherman (1893) who examined the average sentence-length in texts from different eras to see how English has changed over time. In particular, Sherman argued that

Sentence-length as indicator of authorship was also studied by the renowned Scottish statistician G. Udny Yule, who inaugurated the modern era of quantitative authorship attribution with his 1939 paper "On Sentence-length as a Statistical Characteristic of Style in Prose." Yule began by examining the consistency of the sentence-length distributions found in the writings of Francis Bacon, Samuel Taylor Coleridge, and Charles Lamb. Finding in all cases "that sentence-length *is* a characteristic of an author's style" (1939: 370), Yule then applied the technique to two actual cases of disputed authorship. The first text Yule attempted to attribute was *De Imitatione Christi*—an anonymous religious treatise published in 1418, which is one of the most important texts in the Christian Canon. Yule compared the sentence-length distribution of the *Imitatione* to the sentence-length distributions of the known works of the Augustan Canon Thomas à Kempis and the Chancellor of the University of Paris Jean Charlier de Gerson—the two main candidates. Overall, Yule found that Kempis' distribution was more similar to the *Imitatione,* and concluded that "these results are completely consonant with the view that Thomas à Kempis was, and Jean Charlier de Gerson was not, the author of the *Imitatio*" (1939:377).[11] Yule then turned to *Observations upon the Bills of Morality*—an early piece of economic writing which is believed to have come from the hand of either John Graunt or Sir William Petty. Based on an analysis of sentence-length, Yule concluded that "the actual authorship of the Observations is not the same as that of the economic writings of Sir William Petty" (1939: 380-81).

Soon after Yule's 1939 publication, the English entomologist C. B. Williams noticed that the positively-skewed sentence-length distributions that Yule had observed in texts (i.e. most sentences will be relatively short but there will also be a long tail of longer sentences) were strikingly similar to the distributions that he had encountered in his research on insects. In a 1940 paper, Williams proposed that these skewed distributions could be logarithmically transformed into more manageable symmetrical normal distributions, as he had transformed the distributions of insect species. Williams

mean sentence-length has been steadily decreasing for the past 500 years. While Sherman was not directly concerned with authorship attribution, many of his observation are relevant: e.g. in *Analytics of Literature* (1893:256-268) Sherman notes that Macaulay's works exhibit a remarkably stable average sentence-length of 23.5 words per sentence. For criticisms of Sherman's techniques see Moritz (1903).

[11] Gerson had a mean 23 words/sentence and a median of 19 words/sentence, and Kempis had a mean of 18 words/sentence and a median of 15 words/sentence, whereas *De Imitatione Christi* had a mean of 16 words/sentence and a median of 14 words/sentence.

tested his lognormal technique on works by G. K. Chesterton, H. G. Wells, and G. B. Shaw, but while Williams found that these three authors were characterized by consistent, unique and symmetric lognormal sentence-length distributions, he warned that it was still too early to accept sentence-length as a universal indicator of authorship.[12] However, other investigators have questioned the usefulness of the Williams' technique: Holmes (1985) reports that most authors' sentence-length distributions turn out to be negatively-skewed after a logarithmic transformation.

In 1957, William C. Wake used Yule's sentence-length method, with Williams' lognormal adjustment, to set (1957: 331)

> the study of sentence-length distributions of Greek classics on a firm foundation so that they may be used by scholars where appropriate, as part of their standard equipment for examining works of unknown or disputed authorship.

To this end, Wake determined the sentence-length distributions for a selection of works by Plato, Aristotle and Xenophon, all of which he found to be fairly stable—though the similarity of Xenophon's and Aristotle's sentence-length distributions is undeniable. Wake then used this data to attribute two ancient texts of famously uncertain provenance: Plato's *Seventh Letter* and the shared books of Aristotle's *Ethics.* In the first case, Wake found that the sentence-length distributions of the *Seventh Letter* and Plato's known works "shows an agreement so good that statistical proofs of the insignificance of the differences are superfluous" (1957:243). The second case involved the intersection of the *Nicomachean* and *Eudemian Ethics*: while the *Nicomachean Ethics* are considered to be the work of Aristotle and the *Eudemian Ethics* are considered to be the work of one of his pupils, the two collections share three books, which are thus of uncertain authorship. Wake set out to compare their sentence-length distributions in order to determine from which of the *Ethics* the shared books originated. But Wake's results posed more problems than they solved. For while Wake found consistency across the shared texts, and while Wake found that the shared texts resembled samples of Aristotle's known works, Wake also found that the shared texts did not resemble the unshared texts of either the *Eudemian* or the *Nicomachean Ethics*. Furthermore, Wake found that the unshared texts of the *Nicomachean Ethics* were not even consistent in and of themselves. Because of this variation, Wake was forced to conclude that there may have been more than one

[12] See also C. B. Williams (1970:52-63), for more results and data using sentence-length distributions.

author of the *Nicomachean Ethics*—the very texts that he had assumed at the outset were written by Aristotle. Wake accepted this conclusion but we should not: when an attribution study comes to a conclusion that contradicts its original assumption, it may be that it is the method and not the assumption that is flawed.

In 1965, the Reverend Andrew Q. Morton—perhaps the most prolific investigator of authorship—also used sentence-length in his attempt to resolve the problem of the Pauline Epistles.[13] Morton began by testing the stability of sentence-length distributions in the works of the Greek authors Herodotus, Thucydides, Lysias, Demosthenes, and Isocrates. Discovering an adequate degree of stability and uniqueness across these authors' texts, he then applied his technique to the Pauline Epistles. Based on a comparison of the Epistles' sentence-length distributions using the Chi-square test, Morton concluded that only *Romans*, *I Corinthians* and *Galatians* and all but the first fifty sentences of *II Corinthians* were written by the same author.

In the ensuing discussion, the members of the Royal Statistical Society, to whom Morton had presented his paper, were, for the most part, thrilled with the Reverend's research. As I. J. Good (1965:227) wrote in his brief discussion piece: "I have very much pleasure in seconding the vote of thanks for an unusual and stimulating paper." Then again, inspired by Morton's blend of theology and mathematics, Good also took time to explain in his brief reply that (1965:225):

> the ultra-intelligent beings that are almost certainly scattered throughout the universe might well be in telepathic communication with one another, and, if they are, I believe they would constitute the cells of an immortal consciousness that could be called God.

A more useful, if sobering, assessment of Morton's study was provided by the quantitative linguist Gustav Herdan (1965a). Herdan began by attacking Morton's lack of scholarship, noting, for example, that Morton's claim that "the first successful attempt to establish a statistical indicator of authorship was made by Dr W. C. Wake in 1957" (1965a:169) can be refuted simply by reading the introduction to Wake's paper, where the work of both Yule and Williams is acknowledged. Herdan then refers Morton to his

[13] Two other methods used by Morton in this (1965) paper will be considered in sections 3.11 & 3.12 below. Morton also uses sentence-length and/or returns to the problem of the Pauline Epistles in a number of studies, e.g. Levison & Morton & Wake (1966), Morton & McLeman (1966), Michaelson & Morton & Wake (1978) and Morton (1978). See section 2.12 for a detailed review of Morton's research.

own research in *Type Token Mathematics* (1960), where he had demonstrated that there exists far too much within-author sentence-length variation for the measurement to be proffered as a general indicator of authorship. As Herdan wrote, sentence-length is perhaps “the most unsuitable of all available tests” (1965a:229). Herdan’s conclusion is clear: “over 20 pages of irrelevant analysis of classical Greek prose are used to preface the rather disappointing results for the Paulines” (1965:231).

Sentence-length as an indicator of authorship has not fared well since Herdan’s damning critique. Many other investigators of authorship have also rejected this measurement before they moved on to develop new and more successful techniques. For example, Frederick Mosteller found that the average sentence-lengths of Alexander Hamilton’s and James Madison’s writings were far too similar (mean: 34.55, 34.59; standard deviation: 19.2, 20.3) to be of any use in distinguishing between these two authors’ texts (Mosteller & Wallace 1980:6-7). And Alvar Ellegård reported that sentence-length is of little use in a case of disputed authorship involving a large number of possible authors because “the variability within each author largely overlapped the variability between authors” (1962:10).[14]

But the technique has been applied with some success and justification when used in conjunction with other attribution tests. For example, in a notable series of authorship studies, Geir Kjetsaa (1978, 1979) and his team of Scandinavian experts used sentence-length, and other techniques, to attribute *And Quiet Flows the Don* to Mikhail Sholokhov—the Russian Nobel laureate under whose name the novel was originally published. His authorship of the text had been challenged in a 1974 article by a critic known only as D* who claimed that the novel had been written by Fyodor Kryukov—a Cossack novelist who had served in the Don Army and was known to have been writing a novel about the Don Cossacks when he died of typhus in 1920 as they retreated from the communist hordes. The manuscript was then supposedly obtained by Sholokhov who scrubbed out its anti-Bolshevik themes before publishing it under his own name. Kjetsaa and his team determined that D*’s allegations were unfounded: all the tests, including those based on sentence-length, concluded that Sholokhov was the novel’s sole author.

[14] See also K.R. Buch (1952), M.P. Brown (1963), Ellison (1965), O’Donnell (1966), for more discussion of the problems with sentence-length as an indicator of authorship.

Yehuda T. Radday also used sentence-length, in addition to a number of other indicators of authorship, in his attribution of the Book of Isaiah. But Radday had little confidence in this textual measurement, warning that while (1970:69)

> this test turned out to be of a very high indicative value...it was open to the objection that it was I and not Isaiah who decided where a sentence ended.

This problem with sentence-length as an indicator of ancient authorship has been noted by Yule (1939) and Herdan (1965a) and Kenny (1986) and even Wake, the main proponent of sentence-length as an indicator of authorship, who admitted that (1957:334)

> it is usually contended that punctuation marks are a relatively modern invention and even if modern editors agree to their placing, there is no guarantee that this corresponds to the author's original suggestion.

However, Wake concluded that he was justified in using the technique because ancient texts can generally be split into sentences based on an analysis of the text's meaning, grammar and spacing. It seems to me that Wake, and others who followed, including Morton, were justified in making this assumption: in essence, they were analyzing a syntactically-annotated corpus, which is a common and legitimate act of textual preparation, as long as it is done in a sensible and consistent manner.

2.5 PUNCTUATION

The first published study to measure a text's punctuation mark frequency was W. B. Smith's 1888 investigation of the consistency of the Pauline Epistles. In this study, Smith considered the relative frequency of periods, question marks and colons per page, in addition to average sentence-length and the relative frequency of function words (see sections 2.4 and 2.11). Based on these tests, Smith concluded that only *Romans, I and II Corinthians,* and *Galatians* were written by Saint Paul.

In a more recent study, Bernard O'Donnell (1966) looked at the values of twenty-three textual measurements including sentence-length, word-length and the relative frequency of semicolons and dashes in his investigation of Stephen Crane's unfinished novel *The O'Ruddy*. This novel constitutes a unique case of disputed authorship because its first twenty-four chapters were supposedly finished when Crane died in 1900, but the novel was completed three years later by the novelist Robert Barr. After the thirty-two-chapter novel was published, Barr then claimed that he had in fact written three-quarters of the text and demanded that Crane's widow adjust his cut of the profits accordingly. O'Donnell examined the text to see if could detect any quantitative shift in the text's style, and found that Barr had probably been attempting to bilk the widow: the first twenty-four chapters appeared to have been written by Crane, the twenty-fifth chapter appeared to have been written by both authors, and the last seven chapters appeared to have been written by Barr.

The forensic scientist Carole E. Chaski also looked at punctuation mark frequency in her 2001 comparison of a number of attribution methods. In addition to using the Chi-square test to compare the frequencies of common punctuation marks, she also proposed an interesting method based on syntactically classified punctuation marks, where punctuation marks of all types are counted as either being Sentential, Clausal, Phrasal, Appositive or Word Internal. Unfortunately Chaski's experiment was so small, involving only one "anonymous" test text, that one cannot draw any conclusions about the attribution potential of the various measurements.[15]

Overall, there have therefore been surprisingly few attribution studies that have directly examined a text's punctuation. This lack of interest in punctuation mark frequency as an indicator of authorship is probably tied to the general rejection of sentence-length as an indicator of authorship. But if one is attributing modern texts, where punctuation is far more often the product of the author, and where there is a great deal of optionality in how an author chooses to use these grammatical characters, it would seem that punctuation could be a powerful indicator of authorship.

2.6 CONTRACTIONS

The tradition of statistical research in Shakespearean Canonical studies, which had been established by the New Shakespeare Society, was carried into the early twentieth century by such researchers as Ashley H. Thorndike and Willard E. Farnham. These literary scholars focused on the frequency of contractions. Thorndike introduced the method in his 1901 book, *The Influence of Beaumont and Fletcher on Shakespeare*, where he used this technique to support his thesis that Fletcher and Shakespeare were coauthors of

[15] See Grant & Baker (2001) for an article-length criticism of Chaski's methods and results.

Henry VIII. In particular, Thorndike looked at the fluctuations in the proportion of *them* to *'em* in different parts of the text.

Thorndike's method was then expanded by Farnham in an impressive 1916 paper, "Colloquial Contractions in Beaumont, Fletcher, Massinger, and Shakespeare as a Test of Authorship," in which he examined the frequencies of a whole range of contractions in the works of various Elizabethan authors. Specifically, Farnham identified three types of common *colloquial* contraction (so-called because contractions were and still are a feature of non-standard English): *t-contractions* (e.g. *in't* vs. *in it*, *pour't vs. pour it*), *the-contractions* (e.g. *i'the vs. in the*, *o'the vs. of the*), and *s-contractions* (e.g. *on's vs. on his*, *on's vs. on us*). In order to compare texts, Farnham counted the occurrences of each of these contraction-types and then arranged the results in "tables of linguistic preference" so that the rates of contraction in the texts could be compared. After Farnham had tested the technique by generating tables of linguistic preference for samples of the known works of the four authors in the title of his paper, he then applied his technique to cases of disputed Elizabethan authorship.

First, Farnham examined the use of contractions in *The Captain*, a play that was once performed for King James, and which was generally thought to have been written by Fletcher and an unknown collaborator—Beaumont was considered to be the most likely candidate. Farnham found that the first four acts of *The Captain* averaged 1 contraction every 36, 63, 40, and 69 lines, but only 1 s-contraction every 1122 lines over all four acts. Based on this evidence he concluded that these acts were sufficiently consistent with Fletcher's known contraction rates (a relatively colloquial average of 1 contraction every 39 lines, but only 1 s-contraction every 862 lines) for these acts to be attributed to Fletcher. But Farnham found that the disputed fifth act was far too colloquial to be attributed to Fletcher (averaging 1 contraction every 18 lines and 1 s-contractions every 88 lines). However, it did not appear to be the work of Beaumont either, who averaged only 1 contraction per 111 lines of play. In search of a new collaborator, Farnham considered a series of other possible authors, including Middleton, Dekker, Heywood, Day, Ford, and Brome but "none seemed to be the man" (1916:338). Then Farnham happened upon the work of Robert Davenport, "in every way a minor dramatic poet" (1916:338), whose use of contractions was similar enough to *The Captain*'s fifth act (Davenport averaged 1 contraction every 31 lines and used s-contraction at a rate of 1 s-contraction every 253 lines) for Farnham to declare that this obscure writer was the second author of *The Captain*. It is also worth noting, though Farnham did not, that his table of linguistic preferences for Shakespeare's *A Winter's Tale*, which was written around the same time as *The Captain*, displays an average rate of 1 contraction every 20 lines, and 1 s-contraction every 189 lines, both of which are far more similar to the rates found in the fifth act of *The Captain* than are the work of Davenport. Perhaps Shakespeare was Fletcher's unknown collaborator. Farnham also applied his technique to the non-Fletcherian parts of *Henry VIII* and *The Two Noble Kinsmen*, and confirmed the standard attribution of the disputed scenes to Shakespeare over Philip Massinger.[16]

2.7 Vocabulary Richness[17]

Measures of vocabulary richness are often used in authorship attribution. Investigators who use these techniques assume that the frequencies of words in a text are primarily a function of its author's vocabulary. As David Holmes writes (1985:334):

> The basic assumption is that the writer has available a certain stock of words, some of which he/she may favour more than others. If we sample a text produced by that writer, we might expect the extent of his/her vocabulary to be reflected in the sample frequency profile. If, furthermore, we can find a single measure which is a function of all the vocabulary frequencies and which adequately characterizes the sample frequency distribution we may then use that measure for comparative purposes.

There is an obvious problem with this assumption of randomness: the vocabulary of a text depends far more on its subject than on its author. For while every word in a text must be drawn from its author's vocabulary, different subjects will activate different sections of an author's vocabulary, and different sections of any author's vocabulary will not all be equally rich. For example, an author who writes a five-hundred word text about a subject of which he is very knowledgeable will probably use a greater number of word-types than if that text was about a subject of which he has very little knowledge. Furthermore, even if the two texts discuss the same subject, one would expect that a

[16] For other contraction-based attributions, see Elliott & Valenza (1996), where contraction frequency is one of many textual variables examined in their analysis of the Shakespearean canon.

[17] I include measures of vocabulary richness and diversity under one heading. For general references, see Holmes (1985), Baayen (2001), Johnson (1973), Williams (1970), Tweedie & Baayen (1998).

general review of the subject would exhibit a higher vocabulary richness than a text that concentrates on one specific issue. But while this fundamental assumption of any measurement of vocabulary richness is clearly flawed, it may still be a useful assumption to make in quantitative authorship attribution, especially in cases with a limited number of possible authors and where the subject of the input texts can be controlled.

There are two basic textual measurements that are included in any formula for vocabulary richness: N (the length of the text or the number of word-tokens), and V (the size of the text's vocabulary or the number of word-types). Of course, if one wants to compare the vocabulary richness of texts of different length, then one cannot compare the values of V directly because the number of word-types in a text depends on the length of a text. Unfortunately, the obvious solution of taking either

(1) *Mean Word Frequency (MWF)* = N / V

(2) *Type-Token Ratio (TTR)* = V / N

as a measurement of vocabulary richness is also affected by changes in sample size. This is because while the value of V will stabilize as texts get longer and longer, the value of N will continue to grow. For example, if the TTR is calculated for different length samples of a single text, one does not find that its value converges on a population mean with an increase in sample size; rather, one finds that its value will continue to fall as more yet to be seen word-types are introduced into the text. As Harald Baayen writes (2001:1)

> This property sets lexical statistics apart from most other areas in statistics, where an increase in the sample size leads to enhanced accuracy and not to systematic changes in basic measures and parameters.

The main goal of vocabulary richness research has therefore been to devise a new measurement whose value is independent of sample size.[18]

In an attempt to achieve this goal, many measurements have been proposed that are not only a function of a text's word-token and word-type counts but also of a text's *grouped word frequency distribution*—i.e. the number or word-types that occur once in text, twice in a text, etc.. Generally, a word frequency distribution will begin at its highest point—because in most texts the once occurring word-types, or the *hapax legomena,* are the most common—and then fall off, decreasing quickly at first before leveling out with a

[18] Despite this problem TTR & MWF are still used in authorship attribution, see for example Baker (1988), Whissell (1996), Martindale and McKenzie (1995). Also Ledger (1995) uses TTR, though he is careful to only analyze and compare texts of the same size.

long tail of function words (e.g. *the, a, of, and, to*) that occur very frequently in the text but do not share their token frequency with any other word-types. In this paper, I will denote the number of word-types which occur *i*-times in a text by the symbol V_i. For example, V_1 refers to the number of *hapax legomena.*

In his classic 1944 book *The Statistical Study of Literary Vocabulary*, George Udny Yule proposed the first of these more complex measures of vocabulary richness. Yule called his measurement *The Characteristic*[19]:

$$(3) \qquad K \quad = \quad 10^4 \left(\sum i^2 V_i - N \right) / N^2$$

In essence, *K* is a measurement of word repetition rate: the faster that an arbitrarily chosen word in a text is likely to repeat itself, the larger the value of *K*. Therefore *K* is actually an inverse measurement of lexical richness: the larger a text's value of *K* the lower the text's vocabulary richness.[20] After Yule had tested his measurement's consistency and uniqueness across the nominal vocabularies[21] of a number of different authors and a number of different text-lengths, he was confident enough to apply his technique to the problem of *De Imitatione Christi.* Yule began by demonstrating that there was sufficient consistency and uniqueness in values of *K* across the nominal vocabularies of the known works of Kempis and Gerson to discriminate their writings. Yule then calculated the value of *K* for the *Imitatione* and compared it to the *K*-values of Kempis and Gerson's known works, and found that Kempis was the most likely author of the *Imitatione.*

While Yule's Characteristic was innovative it was not perfect. For example, Ellegård (1962) found that *K* was unable to distinguish between a large set of possible authors, and Tallentire (1972) found that the value of *K* varied widely across texts of a single author. Other experts, less concerned with the technique's potential to resolve cases of disputed authorship, took issue with Yule's claim that the measurement was independent of sample size. These criticisms led to a wave of new measures of vocabulary richness, some of which are provided below:[22]

[19] Where 10^4 is an arbitrary constant used to avoid small and difficult to read *K*-values.

[20] C. B. Williams (1970) suggested that this insignificant problem should be resolved by taking the reciprocal of *K*, a measure which Williams calls *Yule's index of diversity.*

[21] Yule focused on nouns in order to limit the number of words he had to count by hand, and because he believed that they would be better indicators than other word categories, especially function words.

[22] For a more detailed review see Baayen (2001:24-32), Tweedie & Baayen (1998), Holmes (1985).

(4)	D	=	$\sum V_i (i / N) ((i-1) / (N-1))$	Simpson (1949)
(5)	R	=	$V / \sqrt{N}$	Guiraud (1954)
(6)	C	=	$\log V / \log N$	Herdan (1960. 1964)
(7)	H	=	$(100 \log N) / (1 - V_1/V)$	Honoré (1979)
(8)	S	=	V_2 / V	Sichel (1975)
(9)	k	=	$\log V / \log (\log N)$	Dugast (1979)
(10)	a^2	=	$(\log N - \log V) / \log^2 N$	Maas (1972)
(11)	U	=	$\log^2 N / (\log N - \log V)$	Dugast (1979)
(12)	LN	=	$(1 - V^2) / (V^2 \log N)$	Tuldava (1977)
(13)	M	=	V / V_2	Michéa (1969, 1971)
(14)	H	=	$-100 \sum p_i \log p_i$	Holmes (1985)[23]
(15)	W	=	$N^{V \wedge -a}$	Brunet (1978)[24]

However, in their extensive review of the various measurements of vocabulary richness, Tweedie and Baayen report that only *K* and *D* are theoretically constant, while all other measurements "may reveal significant deviation from their expected values in actual texts" (1998:334).

Nonetheless, there have been some recent developments in the use of measurements of vocabulary richness in attribution research. In the most significant of these studies, David Holmes (1992) used cluster analysis and principal components analysis[25] to compare the values of five vocabulary richness measurements (K, R, S, α and θ)[26] in his investigation of the texts of the Mormon Canon. The most important text in the Mormon Canon is *The Book of Mormon*, which was supposedly written by six Jewish prophets living in America in the sixth century B.C.. If the text were ever known to ancient man, it was lost to modern man until it was discovered by Joseph Smith in 1827, engraved on golden plates that were hidden in the hills of New York State. According to Mormon tradition, Smith then translated the *Book of Mormon* into Modern English with the aid of magical seer stones. After the ancient texts were translated, angels are then said to have descended from heaven to retrieve the golden plates; but, armed

[23] Where p_i is the probability of the *i*th word-type. This is a modified form of the equation for entropy.

[24] Where *a* is a constant, usually set between 0.165 and 0.172, so that *W* is constant and independent of *N*. See Baayen *et al.* (1996), Holmes & Forsyth (1995).

[25] See section 2.11 for more on the use of advanced multivariate statistics in authorship attribution.

[26] Where α is the slope of the head and θ is the tail of the Sichel distribution. See Sichel (1975), Holmes & Forsyth (1995), Holmes (1996), and Pollatschek & Radday (1981, 1985), for more on its use in attribution studies. See Tweedie & Baayen (1998), for a criticism of use of Sichel Model in attribution studies

with his translations, Smith went on to found the Church of Jesus Christ of the Latter-day Saints. Holmes also investigated two other works of Mormon scripture: the *Doctrines and Covenants* (a series of revelations from Jesus Christ supposedly revealed to Smith) and Smith's "translation" of *The Book Abraham* (a 4000 year old papyrus scroll supposedly acquired by Smith in Kirtland, Ohio). Holmes compared all three of these Mormon texts to samples of Smith's known writings and to the Book of Isaiah, as a control. The goal of the study was thus to discover if Smith's writings would cluster with the Mormon texts, and whether the texts of the individual prophets would form individual clusters. Based on his multivariate vocabulary richness technique, Holmes found that while all of the texts from the Mormon Cannon clustered together as if they had been written by a single author, they did not cluster with the writings of Joseph Smith, which formed a distinct cluster of their own. Holmes' explanation is that "the *Book of Mormon* sprang from the prophetic voice of Joseph Smith himself" (1992:118) and that therefore one should not expect the Mormon texts to cluster with Smith's personal writings. While this may be true, Holmes offers no evidence that this is the case: he does not analyze any texts that are known to have been written by Joseph Smith in his prophetic voice. Indeed, the samples from *Isaiah*, which also form a distinct cluster, are closer to the Mormon cluster than the non-prophetic texts of Joseph Smith. So while it would appear that the works of Mormon scripture were the product of a single author, and not of numerous ancient prophets, there is no conclusive evidence that Smith is their author: perhaps he was presenting another man's work, such as the Clergyman Solomon Spaulding, who has also been considered to be a possible author of the *Book of Mormon.*[27]

2.8 GRAPHEMES

The second new indicator of authorship that was introduced by Yule in *The Statistical Study of Literary Vocabulary* was the frequencies of the letters of the alphabet. Yule discovered the discriminatory value of this measurement by chance during his study of

[27] Holmes & Forsyth (1995) also uses Holmes' multivariate vocabulary richness technique in their study of the *Federalist*; and Baayen *et al.* (1996) use it to set a baseline against which to compare their rewrite-rule-based attribution algorithm. For other attribution studies that use vocabulary richness see Radday (1970), Kjetsaa (1978), Chaski (2001), Pollatschek & Radday (1981, 1985), Somers & Tweedie (2003), P.E. Bennett (1957).

vocabulary richness. When conducting his experiments, Yule had recorded the number of times that each noun-type occurred in an author's corpus on individual filing cards—one card for each noun. Yule stored these cards in a chest of drawers, where each drawer contained a series of cards sorted in alphabetical order representing a particular author's nominal vocabulary. One day, Yule had the adjacent Bunyan and Macaulay drawers open at the same time and noticed that while the two drawers contained roughly the same number of cards, the alphabetical dividers did not line up: the two authors seemed to prefer words beginning with different letters of the alphabet. Yule investigated further and found that the differences were largely a result of Bunyan's preference for nouns beginning with the graphemes *B, F, H* and *W*, and Macaulay's preference for nouns beginning with the graphemes *A, I* and E. As Yule wrote (1944:189)

> Samples from one and the same author do show a degree of consistence notably greater than samples from different authors; so the distribution appears to be definitely characteristic of the *author*.

Yule's explanation for this pattern was insightful: he found that Bunyan used more nouns with Germanic origins and fewer nouns with Latin origins than Macaulay, and that nouns that begin with the letters *B, F, H* and *W* tend to have Germanic origins, whereas nouns that begin in *A, I* and E tend to have Romance origins. Yule saw this property of literary texts as a product of our language's unique heritage (1944:204):

> It is an interesting fact that, owing to the existence of these two great separate streams in our language, even the alphabetic distributions of an author's vocabulary may be distinctive and characteristic.

While Yule was intrigued by this result, aside from Gustav Herdan (e.g. 1966a), investigators of authorship had shown little interest in this novel technique, until Thomas Merriam rediscovered grapheme frequency as an indicator of authorship, in a series of attribution studies.

In the first of these papers, Merriam (1988) examined the frequency of graphemes in an attempt to determine if Thomas Heywood had written the addition to *The Book of Sir Thomas Moore,* which had been added to the original text in a second hand. Merriam's rational for using this technique was (1988:455)

> [Because] the pertinent text samples are old-spelling texts, with the proverbial latitude in spelling before English written conventions were standardized, it is reasonable to examine the letter frequencies of the text.

To test his theory, Merriam compared the grapheme frequencies of a series of texts written by Heywood and his contemporary Anthony Munday. Merriam found that despite the general similarity of the grapheme frequencies in the two authors' samples, there was an even greater degree of similarity across the works of the individual authors. Based on this result, Merriam felt justified in setting up a threshold test: when the grapheme frequencies of the mysterious addition to *The Book of Sir Thomas Moore* is compared to grapheme frequencies of Heywood's known works, if the resultant correlation coefficient is above .990, then the texts are by Heywood. Based on this test Merriam concluded that the addition was not from the hand of Heywood.

Merriam also analyzed the grapheme frequencies of texts with standardized spelling, in a 1994 authorship study using samples of modern editions of Marlowe's and Shakespeare's works. While Merriam did find more similarity overall between the standardized Marlowe and Shakespeare texts than between those non-standardized Heywood and Munday samples that he had analyzed in 1988, he also found that the relative frequency of the letter *O* seemed to allow him to distinguish between these two authors: all 36 of Shakespeare's plays had a relative frequency of *O* over 0.078 and 6 out of 7 plays by Marlowe had a relative frequency of *O* under 0.078. In conclusion, Merriam offered the following amended explanation of grapheme frequency as an indicator of authorship (1994:469):

> Differences in letter frequencies in modern spelling either reflect personal differences in the use of common words such as 'no' and 'not', or reveal deeper registers of phonetic explanation.

Of course, other explanations, such as Yule's etymological rationale, are possible. Or perhaps Shakespeare simply preferred names that contained an *O* (e.g. Romeo, Othello, Antony and Cleopatra).

In a second 1994 study, Ledger and Merriam examined *The Two Noble Kinsmen*, using more complicated multivariate techniques (cluster analysis, principal components analysis) to compare the grapheme frequencies in known samples of Shakespeare's and Fletcher's works to those found in *The Two Noble Kinsmen*. Merriam and Ledger's results for an act-by-act attribution agreed with the standard literary analysis of mixed Shakespeare-Fletcher authorship. Finally in a 1998 paper, Merriam used neural networks, where the input nodes were activated by occurrences of the letters *A, N, O, R* and *Y,* to investigate the Shakespeare-Marlowe *Henry V* case of disputed authorship.[28]

In another recent study, Gerard Ledger (1995) used a set of more complex grapheme-based measurements in his investigation of the Pauline Epistles. In particular, Ledger used two advanced multivariate data analysis techniques (canonical discriminant analysis and principal components analysis) to analyze the percentage of words that contain at least one instance of a specific grapheme and the percentage of words that end in a specific grapheme.

Despite Yule and Merriam and Ledger's success using grapheme frequency as an indicator of authorship, the technique is still usually disregarded by modern investigators and is even criticized. For example, in his review of the field of authorship attribution, David Love attacked investigators of authorship for not providing a rationale for the particular measurements that they use, and cites Merriam's 1994 use of grapheme frequency as a prime example (2002:159-60):

> If we did know why, it might be possible to hone in on the real basis of discrimination rather than counting the letters of the alphabet. This, I take it, is precisely what a scientist would want to do.

While Love is certainly right to insist that investigators consider the rationale for their measurements, there is nothing unscientific about counting the frequency of graphemes. Indeed, the attribution potential of grapheme frequencies has also been verified independently by Forsyth and Holmes (1996), though they found that grapheme frequencies were the least successful of the textual measurements that they evaluated.

2.9 ETYMOLOGY

In his 1887 response to Mendenhall's original article, in addition to sentence-length, H. T. Eddy proposed that authors could be distinguished by comparing "the character of the vocabulary as regards to derivation from Anglo-Saxon, French, Latin, Greek, etc." (1887:297). This is the very same factor that Yule had used to justify his word-initial grapheme method, but if the goal is to develop an etymological-based attribution method, then a far more direct approach is to count word origins.

[28] See section 2.11 below for more references on connectionist approaches to authorship attribution.

This was the approach taken by Gustav Herdan in his 1956 attribution study of *The Equatorie of the Planetis*—a Middle English astronomical manuscript discovered in a Cantabrigian library, which appeared to be the long-lost chapter of Chaucer's *Treatise on the Astrolabe.* Chaucerian authorship was supported by an earlier analysis of the text's rare words, many of which were known to have occurred, in Middle English, only in the works of Chaucer. Herdan set out to uncover additional evidence of Chaucerian authorship through an analysis of the text's Romance vocabulary. While Herdan found that Chaucer's use of Romance words seemed to distinguish him from other writers, the percentage of Romance vocabulary in Chaucer's texts was also dependent on sample size. To overcome this complication, Herdan introduced a formula for calculating the percentage of Romance vocabulary in a text by Chaucer, which he claimed was independent of sample size:

$$(16) \qquad P_V \quad = \quad 10\, log_{10} N$$

The *Equatorie* contains approximately 6048 words, and so Herdan estimated that the percentage of words with Romance origins in the *Equatorie,* if it were written by Chaucer, would be 37.8 %. Indeed, 37% of the text's vocabulary was of Latin origins, and so Herdan concluded that (1956: 258):

> The agreement between observation and expectation, or between fact and theory, is so striking that without going further into the question of statistical significance we may conclude that by the token of Romance vocabulary the *Equatorie* is to be regarded as a work by Chaucer.

Unfortunately, there has been little interest in applying and testing Herdan's method on a wider range of problems.

2.10 ERRORS

In his review of internal-evidence-based forensic authorship attribution, Gerald McMenamin (1993: chap. 6) reintroduced the work of the early forensic document examiners, including Albert S. Osborn, whose Questioned Documents (1910, 1929) was the first major treatise on forensic authorship attribution. Here, Osborn provided an inventory of internal indicators of authorship, which he insightfully divides into three categories, which he calls "Subject Matter," "Rhetoric, Composition, Language" and "Errors." While most of the indicators he lists in the first two categories are far too subjective to be quantified in any straight forward way (e.g. “hysteria,” “incoherence,” “insanity”), his list of error-types is instructive, and most could be readily quantified:

> (1) Spelling; (2) Capitals; (3) Punctuation; (4) Paragraphing; (5) Titles; (6) Person; (7) Number; (8) Case; (9) Pronoun and antecedent; (10) Verb and subject; (11) Mood; (12) Tense; (13) Voice; (14) Possessives; (15) Omissions; (16) Interlineations; (17) Erasures; (18) Repetitions; (19) Facts or statements.

Osborn’s research was influential, and the analysis of errors as a means of identifying authorship is now an important component of forensic document examination.[29]

More recent studies of digital texts have also used error- and format-based measurements to resolve cases of disputed authorship. For example, using a corpus of newsgroup postings, Koppel and Schler (2003) tested a method based on the values of a number of error-based textual measurements, including the frequency of sentence fragments, run-on sentences, repeated words, missing words, plurality and tense agreement errors, missing hyphens, *that*s following commas, and various types of spelling errors (confused letters, repeated letters, etc.). But Koppel and Schler failed to specify what constitutes a run-on sentence, how to determine if a word is missing from a sentence, how to distinguish between grammatical errors and grammatical variation, and they do not acknowledge, as I now demonstrate, that the word *that* can follow a comma, no matter what their grammar-checker may flag. These examples illustrate a basic problem with error-based attributions: the identification of errors is best accomplished through a subjective reading of a text. As every modern writer knows the spell-checker is not always right. Furthermore, it is questionable whether an intelligent reader can even identify an error: if one defines an error as any string of characters in a text that the author did not intend to write, then only the writer can identify errors. Perhaps the author purposely misspelled a word or used non-standard syntax to make a joke. Despite this basic problem with error-based attributions, Koppel and Schler found that their technique was more accurate than a word frequency technique and a part-of-speech frequency technique that they also tested.[30]

[29] Errors analysis was also endorsed in forensic texts such as Conway’s *Evidential Documents* (1959) and Hilton’s *Scientific Examination of Questioned Documents* (1982).

[30] None of the techniques performed particularly well, probably because the texts are quite short averaging only two-hundred words per message. On longer texts, and text with a more standardized format than internet messages, these results would probably not hold. For an attribution study of emails see de Vel *et al.* (2001).

Another reason that attributions should not usually be based on an analysis of a text's errors is because there is no reason to believe that most spelling and grammar errors will be consistent across an author's works: many mistakes are likely to be corrected over time. Don Foster, the well-known authorship investigator who uses a mixture of qualitative and quantitative techniques, offers a particularly good example of why spelling errors are unreliable indicators of authorship. In *Author Unknown* (2000), Foster recounts how he had acted as a linguistic expert in the trial of the Unabomber—the anti-technology terrorist who had been sending mail bombs to engineers and capitalists. Ted Kaczynski, a disgruntled math professor living in the wilds of Montana, had been identified as the Unabomber when his family recognized his style and ideas in the Unabomber's manifesto. The family's attribution was then verified by James Fitzgerald, an FBI expert, which allowed the Agency to arrest Kaczynski at his Montana shack. But during the trial, Fitzgerald's attribution was called into question by the defense and their linguistic expert, the well-known sociolinguist Robin Lakoff, who made reasonable objections to Fitzgerald's reliance on spelling evidence. For example, Fitzgerald's attribution was based in part on Kaczynski and the Unabomber's use of British spellings (e.g. *analyse* and *licence*); but, as Lakoff argues, these spellings are not rare enough in American writing to prove that Kaczynski was the Unabomber. Lakoff also presented spelling-based counter-evidence: for example, the Unabomber had spelt the word *chlorate* correctly, whereas Kaczynski had spelt the word *clorate.* Foster was called in by the prosecution to verify Fitzgerald's attribution. In his testimony he demonstrated the basic problem with error-based authorship attribution: while Kaczynski had never spelt the word *chlorate* correctly, in Kaczynski's newer texts the words *chloride* and *chlorine* do occur and are spelt correctly. Foster argued that if Kaczynski had learnt to spell these words correctly at the time that the Unabomber texts were published, then one should expect that Kaczynski had learnt to spell *chlorate* correctly as well. The judge agreed with the prosecution and Kaczynski is now living in a Colorado penitentiary. [31]

[31] Though Foster uses spelling conventions as evidence in his study of the *Funeral Elegy* (1989).

2.11 Words

The main problem with any univariate attribution method, such as measurements of vocabulary richness, is that there appears to be far too much stylistic and thematic variation across any author's texts, for those texts to be characterized by the value of a single measurement. Many investigators of authorship have thus chosen to consider the values of multiple textual measurements. Today, most multivariate approaches to quantitative authorship attribution are based on the frequencies of individual words.

W. B. Smith was the first investigator to use word frequencies to help resolve a case of disputed authorship. In his 1888 study of the Paulines, Smith considered the frequency per page of thirty-six individual and sets of Greek function words, in addition to measurements of sentence-length and punctuation mark frequency (see sections 2.4 and 2.5). Smith chose to count function words[32] because they (1888a:454)

> Constitute the fibrous tissue of speech, spreading through and permeating the entire organism of discourse...they are the current co-ordinates in the equation of style, determining by their mutual relations the law of form for the writers thoughts.

And to ignore an author's use of content words because

> The stock of words, so far as it consists of notional terms,—verbs, nouns, adjectives, in less degree adverbs,—*must* vary with the subject in hand, and how much variation may be allowed to a given change of subject is a delicate question, an unanswerable one.

Once Smith's had determined the mean values of his forty textual measurements across the four core Pauline Epistles, he graphed their values from most to least frequent to produce what he called an author's *curve of style*. Smith then compared the Pauline curve to the corresponding curves for *Ephesians, Philippians,* and *Colossians*, and concluded that they had not been written by Saint Paul.

Soon after Smith's articles were published they were forgotten, and word frequencies were not reintroduced to attribution studies until Alvar Ellegård published a pair of books in 1962: *A Statistical Method for Determining Authorship* and *Who Was Junius?* Ellegård's goal was to develop a generally applicable statistical method for

[32] Function words are here defined as the members of closed word classes (see Schachter 1985): e.g. in the standard taxonomy of English word classes, function words include the words in all classes except verbs, nouns, adjectives, and adverbs, but including modal verbs, auxiliary verbs and adverbial Particles. See Section 3.4.7 for a more detailed discussion of the distinction between content and function words.

determining the author of an anonymous text, and to then apply this method to the mystery of the anonymous *Junius Letters*—a series of scathing political articles, published anonymously in London's *Public Advisor* from 1769 to 1772, which attacked the morality of Prime Minister Grafton and members of his Whig administration. Ellegård chose *The Junius Letters* because they are a relatively well-known case of disputed authorship and because there is much material upon which to base the attribution: there are 69 *Junius Letters* totaling over 150,000 words, and there are many texts written by the *Letters'* many possible authors, including Sir Philip Francis, the main suspect.

Ellegård used two novel word frequency techniques to compare the *Junius Letters* to the writings of Francis and his contemporaries. Both of these methods were based on the same rationale (1962a:12):

> Some words, phrases, and turns of expression are felt as vaguely "typical" of a particular author, while other words and turns of expression are such as he "would never have used".

The first method involved discovering Junius's characteristic and uncharacteristic words and phrases ("plus words" and "minus words"). To calculate a word's *distinctiveness ratio*, Ellegård divided the relative frequency of the word in the Junius corpus by its relative frequency in a one-million-word control corpus of eighteenth century texts. A word with a distinctiveness ratio over one was considered to be a Junius plus word and a word with a distinctiveness ratio below one was considered to be a Junius minus word. To assemble a set of distinctive Junius words, Ellegård read through the *Junius Letters* and the one-million-word corpus multiple times in search of words that appeared to be either characteristic or uncharacteristic of Junius. Once Ellegård had assembled this initial list of (mainly content) words, he read back through the two corpora manually counting the number of times that each of the words occurred. He then calculated the distinctive ratio of each of these words and kept those that he found to be the most powerful indicators of Junius's prose.[33] The distinctiveness ratios of these words were then compared to their values in each of the candidate's writing samples to determine which author had the most similar lexical taste. Ellegård's second method involved

[33] E.g. Junius plus words: *afflict, engross, callous, indulge, notorious, sordid*; Junius minus words: *apprize, entire, indeed, proper, surely, utmost*.

comparing the frequency of fifty-one sets of synonyms or *proportional pairs* such as *on-upon*, *kind-sort-species*, and *completely-entirely-totally-wholly*, which he had also identified in his readings. Both methods yielded the same result: they agreed with the preponderance of external evidence and attributed the *Junius Letters* to Sir Philip Francis.[34]

Both of Ellegård's techniques were based primarily on a comparison of the frequency of the types of content words that Smith had warned against using to attribute authorship. Ellegård justified his choice based on the theory that (1962a:15-16):

> The words most frequently used in the language—articles, prepositions, conjunctions, and pronouns, as well as the commonest verbs, nouns, adjectives and adverbs—are necessarily about *equally* frequent in texts, whoever the author.

In essence, Ellegård was echoing the position taken by Yule, who wrote (1944:21)

> My object in limiting myself to nouns...was in part simply the limitation of material and the exclusion of words with little or no significance to style, such as prepositions, pronouns, etc.

It was not until the research of Mosteller and Wallace that modern investigators realized, like Smith before them, that content words are not good indicators of authorship, because their frequency depends primarily on the subject and not on the author of a text.

In one of the most influential series of studies in the history of quantitative authorship attribution, the statisticians Mosteller and Wallace (1963, 1964, 1984) used function words to investigate the origins of *The Federalist Papers*—eighty-five political essays published in New York's *The Independent Journal* from 1787 to 1788 in an attempt to persuade the populace of New York State to ratify the Constitution. At the time, it was known that the papers were written by James Madison, Alexander Hamilton and John Jay, but the exact authorship of the individual papers was a mystery because all of the papers were signed with the pseudonyms *Publius* or *A Citizen of New York.* The first paper-by-paper attribution appeared in 1807 when a Philadelphian publication printed a list whose author claimed to have examined Hamilton's own annotated copy of *The Federalist.* Unfortunately the attribution could not be verified: while the executers of Hamilton's will had supposedly donated the text to the New York Public Library, it has

[34] Ellegård's basic method of looking at plus and minus words has been adopted by many other investigators, including Foster (e.g. 1996) and Elliott and Valenza (e.g. 1996) who look at "badge" and "fluke" words, though these investigators use these measurements in slightly different ways.

not been seen in centuries. However, in 1817 this attribution was corroborated when a second list was published, which had allegedly been slipped by Hamilton himself into the bookshelf of a friend only two days before being killed in his famous duel with Aaron Burr. But, in 1818, Madison challenged these lists, after he had left the presidency. The result is twelve *Federalist Papers* that the original lists attributed to Hamilton, but which the fourth president of the United States attributed to himself.

Before settling on frequent words as indicators of Madison's and Hamilton's authorship, Mosteller and Wallace first tested and rejected a number of other textual measurements, including average sentence-length, the frequency of nouns and adjectives, and the frequency of one- and two-letter words. It was in fact Douglass Adair, an expert on the *Federalist Papers*, who had suggested to Mosteller that proportional pairs such as *while/whilst* might be useful indicators of Hamilton's and Madison's authorship. Finding this observation to be true, Mosteller and Wallace looked at the frequencies of 165 frequent words, which were eventually whittled down to the 30 most characteristic words, of which 16 were functional. Mosteller and Wallace were particularly attracted to function words because "many of them are not much influenced by the context of the writing" (1984:17). Using Bayes Theorem to compare the frequencies of these 30 words in Hamilton's and Madison's undisputed papers to those in the disputed papers, Mosteller and Wallace concluded that Madison had written all twelve of the disputed papers. [35]

After Mosteller and Wallace's study was published, many similar studies appeared that examined the origins of Ancient Greek texts. The first of these studies was A. Q. Morton's 1965 investigation of the Pauline Epistles, which was discussed in section 2.4. In addition to sentence-length, Morton examined the frequency distributions of the Greek function word *kai* ("and"), *en* ("in"), *einai* ("to be") and *autos* (a personal pronoun). When Morton compared the frequencies of these words across the Pauline Epistles using the Chi-square test, he found that only the core Epistles clustered together. Anthony Kenny (1978) also looked at the frequencies of a large number of Greek function words (e.g. particles, connectives, prepositions, pronouns) in his study of the

[35] See Francis (1966) for a review of Mosteller and Wallace's study; see Adair (1944a, 1944b) for a review of the *Federalist* problem; and see Kjell (1994), Waugh & Adams & Tweedie. (2000), Hilton & Holmes (1993), Merriam (1989), Farringdon and Morton (1989), Holmes & Forsyth (1995), Tweedie & Singh & Holmes (1996a) for other investigations of the *Federalist*.

shared books of Aristotle's *Ethics*. Using a variety of statistical tests, including Spearman's Rho, Pearson's Correlation Coefficient and the Chi-square test, Kenny compared the frequencies of function words in the shared books to the unique books of the *Eudemian* and *Nicomachean Ethics*, and concluded that the shared books were more similar to the *Eudemian Ethics*.[36]

Since 1988 many attribution studies have analyzed word frequencies using more complex multivariate statistical techniques such as cluster analysis, discriminant analysis, and principal components analysis.[37] These techniques are used in many fields to analyze datasets where samples are represented by the values of a set of variables. In authorship attribution, such cases arise whenever the texts of authors are compared based on the values of multiple textual measurements. The most popular advanced multivariate technique in authorship attribution is principal components analysis (PCA), which is a statistical procedure for transforming the values of a set of measurements into a smaller set of new uncorrelated measurements or principal components (PCs), which are then ordered so that the first two or three PCs account for most of the variation in the dataset (Woods *et al.* 1986). By reducing the number of dimensions that describe the data, a PCA allows for patterns in the dataset to be more easily observed. For example, while it is difficult to make sense out of a dataset that consists of one-hundred text samples each described by the relative frequency of fifty common words, if the dataset can be described by two or three principal components, the investigator can easily construct graphs, with these components as axes, where any clusters in the corpus can be visually identified.[38]

Using PCA to analyze function word frequencies, in particular, was introduced by J. F. Burrows in his innovative and influential 1988 paper, "*Anna Boleyn* and the Authenticity of Fielding's Feminine Narrative." In this study, Burrows used the method to investigate the problem of the authorship of the feminine narratives found in the works of Henry Fielding—the original English novelist. In his novels, Fielding often chose not to tell the stories of his female character in the voice of the text's masculine narrator, but

[36] For more studies of Greek texts using function words see Usher & Najock (1982), Kenny (1986).

[37] Though these multivariate techniques had been used in attribution studies since at least O'Donnell's 1966 study of the *O'Ruddy* (see sections 2.5 & 2.14).

[38] See Wood et al. (1986) for an introduction to PCA in linguistics, and Binongo & Smith (1999a) for introduction to PCA in authorship attribution.

instead to allow these characters to recount their own histories in their own fictional voices. For some time, Fielding experts have wondered if some of these female histories, most especially the narrative of Anna Boleyn in *A Journey*, were not written by Fielding's sister Sarah. This puzzle was brought to Burrows attention by the paper's coauthor, the literary scholar A. J. Hassall, who asked Burrows to compare Anna Boleyn's history to the known writings of the Fielding siblings using the multivariate technique Burrows had used in his book *Computation into Criticism* (1987) to analyze the differences in the dialogue of Jane Austen's fictional characters. Burrows began by subjecting the function word frequencies he found in ten histories by Henry Fielding and ten histories by Sarah Fielding to a principal components analysis, which resulted in a new set of measurements, where the first three principal components accounted for 94.99%, 2.32% and 0.71% of the variation in the dataset. When the values of the twenty undisputed histories were plotted along the first principal component, the works of the two authors were intermingled: there was no discernable pattern in authorship. But when Burrows plotted the histories along the second principal component he found that the works of the two authors formed two remarkably well-defined clusters. When the Anna Boleyn history was included in the dataset and the principal components were recalculated, the disputed history was found to cluster with the histories written by Sarah Fielding. However, upon further investigation, Burrows found that if he split up the history into three sections, then the beginning fell in the Henry Fielding cluster, the middle section fell in the Sarah Fielding cluster, and the final section fell in between the two clusters. Burrows' explanation for this result was that (1988:444):

> Henry Fielding wrote the beginning of the history of Anna Boleyn; that, under the heavy personal pressure of the year 1743 and as one of his many gestures to her literary aspirations, he allowed his sister to continue it for him; and that he either revised her version of the ending or added an ending in which he sought to imitate her style.

Since his pioneering 1988 work, Burrows and his associates at the Centre for Literary and Linguistic Computing at the University of Newcastle, Australia—most notably Hugh Craig—have used the technique in a number of similar attribution studies. For example, in a compelling 1992(a) study, Burrows demonstrated the potential of his technique by using it to distinguish between the writings of the three Bronte sisters. In another study, Burrows and Craig (2001) used the method to verify literary scholar David Norbrook's attribution of two 17th-century poems to Lucy Hutchinson.[39]

One of the first investigators of authorship to recognize the potential of Burrows' technique was M.W.A. Smith, who expressed his interest in his 1990 review of four statistical techniques often used in authorship attribution. Smith then applied the method in a series of short authorship studies (1991, 1992, 1993) with varying degrees of success. In the first study, Smith used Burrows' technique to investigate the problem of *The Revenger's Tragedy*—an anonymous play published in 1607 which experts believe was written by either Cyril Tourneur or Thomas Middleton. Smith found that Middleton's and Tourneur's writings could be distinguished using Burrows' method, and that *The Revenger's Tragedy* clustered with Middleton's works.[40] In the second study, Smith looked at the problem of *Pericles*—a play of disputed Shakespearean authorship which is thought to have been written in part (especially acts I and II) by George Wilkins.[41] Using Burrows' method, Smith compared acts I and II of *Pericles* to acts III through V of *Pericles* to samples of both Shakespeare's and Wilkins' undisputed works and found that both sections of *Pericles* fell disappointingly between Shakespeare's and Wilkins' clusters. In the third study, Smith used Burrows' method to investigate the case of the anonymous play *Edmund Ironside*, which had been attributed by some scholars to Shakespeare. Smith found that the play did not cluster with the works of Shakespeare.[42]

Many other investigators have also applied Burrows' method. For example, in a 1998 study, Tweedie, Holmes and Corns used the method with Latin function words to determine if John Milton wrote *De Doctrina Christiana*—a heretical religious treaty discovered in 1823 wrapped in Milton's personal papers. This case of disputed authorship is significant because if the text were included in Milton's Canon, it would force scholars to thoroughly rethink their understanding of Milton's theology. Tweedie *et al.* compared the *Doctrina* to nineteen samples of Milton's known Latin works, and to texts from eight

[39] See also Craig (1992, 1999), Burrows (1992b) for use of PCA and function words to attribute texts; and see Burrows & Craig (1994) for an example of the method being used to distinguish eras.

[40] Though Smith also compared the writings of Tourneur and Martson and found that both authors clustered together. For more on *Revenger's Tragedy* See Love (2002), Lake, M.W.A. Smith (1987).

[41] For more examples of Smith's studies of *Pericles,* see Smith (1987, 1988, 1989a).

[42] In this paper Smith also argues that Shakespeare may have written *Edward III.* See also Smith & Binongo (1999a) for an intro to PCA in authorship attribution. And the authors also use the method to compare styles of a single author: see Binongo (1993, 1994, 1995); and Binongo & Smith (1999b).

other seventeenth century neo-Latin writers, including eleven samples of Bacon's neo-Latin writings. First, the investigators tested the method on the texts of known authorship and found that the method successfully clustered both the Milton and Bacon texts. When the *Doctrina,* split up into ten samples, was introduced into the analysis, they found that while the samples were more similar to the works of Milton than to any of the other authors, the attribution was by no means conclusive. Furthermore, a fair amount of variation was found across the ten samples of the *Doctrina.* Based on this evidence, the investigators concluded that the "Miltonic status of its constituent parts is uneven and uncertain" (1998:86).[43]

In another recent attribution study, Holmes, Gordon and Wilson (2001) used Burrows' method to resolve a particularly curious case of disputed authorship—the war letters of George Pickett, the great Confederate General who led the daring Pickett Charge which cut deep into Union lines at Gettysburg but left three-quarters of his own men dying in its wake. In 1913, Pickett's widow published the General's Civil War letters, in her best-selling *Heart of a Soldier* (1913). However, soon after the letters were published their authorship was called into question by experts who believed that the letters were the work of the General's wife. To test this theory, Holmes *et al.* compared the letters to a large and well-constructed corpus of texts: the investigators collected letters written by other Civil War soldiers, including Robert E. Lee; samples of Pickett's field reports and letters to other family members; and a large sample of Pickett's wife's own writings. Using Burrows' method, they found that the disputed letters were far more similar to the works of the widow than to the works of the General.[44]

In another study, Thomas Merriam (1996) attributed *Edward III* using a principal components analysis of *word ratios*—a complex textual measurement that combines word frequencies in order to create more potent indicators of authorship. This

[43] See also Campbell *et al.* (1997).

[44] See Holmes and Forsyth (1995); Forsyth and Holmes (1996); Forsyth, Holmes, Tse (1999); Merriam (1998); Holmes, Robertson & Paez (2001), Koppel & Schler (2003), for more applications, extensions and tests of Burrows' method. The publication of Burrows' 1988 paper also led to investigators using advanced multivariate techniques to analyze other types of textual measurements. For example, as already discussed, Holmes (1992) used PCA and cluster analysis to examine multiple measurements of vocabulary richness in his study of Mormon texts (see section 2.7), and Merriam (1998) used principal components analysis to analyze word and grapheme frequencies in his investigation of *Henry V* (see section 2.8).

measurement was introduced by Thomas Merriam and Robert Matthews in a 1993 paper, where word ratios—in particular, *did / (did + do)*, *no / (but + by + for + no + not + so + that + the + to + with)*, *no / (no + not)*, *'to the' / to*, and *upon / (on + upon)*, where each italicized word denotes that word's frequency in a text—were used to investigate the provenance of *The Two Noble Kinsmen.*[45] However, Matthews and Merriam's 1993 paper is even more significant because it introduced neural networks as a method for comparing the values of a set of textual measurements. A neural network is a system of input nodes, hidden nodes and output nodes connected through a series of weighted links. In an attribution neural network, the input nodes are activated by the values of some set of textual measurements, which cause the input nodes to activate the hidden notes in varying degrees, which in turn activate the output nodes in varying degrees, each of which is associated with one possible author. The weighting of links, which determines the degree to which the various nodes will be activated, is set by training the neural network on the possible authors' writing samples. The advantages of using neural networks in authorship attribution is that they are capable of recognizing non-linear patterns in the texts and of classifying the texts based on these patterns even in the midst of noise. Once the network was trained and tested on samples of Fletcher's and Shakespeare's known works, Matthews and Merriam used their connectionist word-ratio technique to attribute the acts of *The Two Noble Kinsmen*: the results agreed with the standard attribution of mixed Shakespearean/Fletcherian authorship.[46]

Since Mathews and Merriam's innovative paper, there have been a number of attribution studies that have used neural networks to analyze the values of sets of textual measurements. For example, Tweedie & Singh & Holmes (1996a) used neural networks to reattribute the *Federalist Papers*, where the input nodes were activated by the

[45] Though Merriam refers to two earlier unpublished papers in which word ratios were used: T.B. Horton (1987) and Merriam (1992).

[46] Merriam and Matthews (1994) conducted a follow up study using similar word ratios and inputs to distinguish between the works of Marlowe and Shakespeare. For more studies using word ratios see also Merriam (1993). For a more complicated method for combining word frequencies, see Holmes & Forsyth (1995), who investigate the *Federalist Papers* using a "Genetic Rule-Finder" that essentially begins by generating a random set of rules consisting of variables (in this case function word frequencies), constants and operators, for example ((ON – THERE) < 2.832), and then goes through a process of evaluating each rule on a training dataset, deleting the least successful rules, and then combining and mutating the remaining rules to generate a new set of rules. The process is then repeated until a set of rules which successfully attributes the writing samples is discovered.

occurrence of those function words discovered by Mosteller & Wallace to be characteristic of Madison's and Hamilton's writings (*any, form, an, may, upon, can, his, do, there, on* and *every*). After training the network on the undisputed papers, Tweedie *et al.* used the network to attribute the disputed *Federalist Papers.* Their results agreed with the conclusions of Mosteller and Wallace.[47]

Despite the great popularity of word frequencies as indicators of authorship—as attested to by the number of methods that have been proposed for comparing their frequencies—these textual measurements have been subject to considerable criticism. Once again, most of this debate revolves around the randomness assumption: can we assume that a word is used at a stable rate across an author's works? Clearly, this is not a valid assumption in the case of content words, whose frequencies are largely determined by the subject matter of a text, but the randomness assumption may hold for function words. In one important study, Fred J. Damereau (1975) argued that most function words are not used in a random manner. Damereau based this claim on an examination of the observed frequency distributions of individual function words: if the word was truly used by an author in a random manner then one would expect that its distribution would approximate the Poisson distribution. This same basic test was conducted by both Mosteller and Wallace (1964) and Morton (1965), and in both cases the investigators found that the distributions of the words being analyzed were adequately similar to the Poisson distribution, though Mosteller and Wallace found that for a number of words the negative binomial distribution provided a better approximation. But, based on an examination of samples of works from Vonnegut, Hemingway, Miller and Thackeray, Damereau discovered that few function words were used in a random manner, and that the small set of words that could be considered to be randomly distributed in the works of one author were unlikely to be randomly distributed in the works of the next. Based on this evidence Damereau concluded that while function words may be useful in the occasional attribution study, they are not good general indicators of authorship.[48]

[47] For more studies using neural networks see Waugh & Adams & Tweedie (2000), Lowe & Matthews (1995) (using Radial Basis Function), Kjell (1994), Merriam (1998), Hoorn (1999), and see also Tweedie & Singh & Holmes (1996b) for a review of neural networks in quantitative authorship attribution.

[48] See also Hoover (2001) for a recent criticism of using Cluster Analysis and frequent word frequencies.

While Damereau's argument has been very influential—often being cited as proof that function word techniques are flawed (e.g. Oakman 1980, Oakes 1998)—it has not been generally accepted. For example, Holmes (1985) quite rightfully questions whether Damereau is justified in making such a strong claim about language in general after considering the work of just four authors: it is quite possible that the stability in function word frequencies observed in the works of Madison and Hamilton are the norm, and that the lack of stability in the works of the four authors that Damereau considered is the exception. Indeed, the strongest piece of evidence in favor of the randomness assumption are the many successful function word-based attribution studies that have been reviewed in this history.

But why would function words be used at stable rates? The standard explanation provided by investigators of authorship is that these words are ubiquitous or that these words are often used unconsciously or that these words express grammatical information. While none of these claims are false, and while all are perhaps necessary for function words to be used at stable rates, they do not explain why function words would be used at stable rates. If one accepts the basic typology of stylistic, thematic and error-based internal-evidence types proposed by Osborn, which in my opinion is correct, then there are three possible basic explanations for the stability in the relative frequency of function words across an author's texts. Both a stylistic and a thematic explanation seem to account for this stability.

The stylistic explanation would be that function word frequencies are determined by the unique variety of language in which an author writes (i.e. function word frequencies are determined by the author's selection between the semantically equivalent variants of a linguistic variables).[49] This is essentially the rationale that Mosteller and Wallace provide when they wrote "we need variables that depend on authors and nothing else" and "some function words come close to the ideal" (1984:266). Now certainly function word frequencies do depend *in part* on stylistic patterns in a text. For example, the frequency of the word *and* must be affected by how an author chooses to conjoin the elements of a list; and the frequency of all pronouns must depend on how often the author chooses to repeat the name of a person or thing that he had already referenced earlier in

[49] See section 4.2.1 for a discussion of linguistic variables and a definition of linguistic style.

the text; and the frequency of the word *not* depends on how often the author chooses to contract the negative marker. But while the frequency of function words is undoubtedly affected by a writer's style, it should be obvious to anyone who has ever uttered a function word that their frequencies depend also, if not mainly, on the meaning that they express. This claim can be verified simply by taking any sentence and interchanging one function word with another: the result will almost always be gibberish or a sentence with a different meaning. In either case there will be a change in the meaning of the utterance. For example, the relative frequency of the word *and* in one paragraph to the next in this review of authorship attribution depends significantly on whether I am reviewing the work of an individual investigator or of a group of investigators. This is not a stylistic pattern. So while function word frequencies may be sensitive to an author's style, their usage, like all other words, depends mainly on the subject of the text.

The fact that function word usage depends largely on the subject of a text does not invalidate the use of function words in quantitative authorship attribution. The randomness assumption may still be a useful approximation, if the type of information that function words convey is the type of information that is expressed by the author at a stable rate. This thematic explanation is not unreasonable: for example, an author who tends to introduce new subjects throughout his texts will tend to have a relatively high number of *the*s; and an author who tends to convey extra information about nouns will tend to have a relatively high number of adjectives interlinked by *and*; and an author who tends to convey information about directions will tend to use a relatively large number of prepositions and adverbial particles. This is the type of general information that an author may choose to convey on a regular basis as a manifestation of his worldview. Function word frequencies may therefore be successful indicators of authorship because they are sensitive to a combination of stylistic and thematic indicators of authorship.[50]

[50] The same is true of almost all the other textual measurements used in quantitative authorship attribution (see section 5.9). This is why I have chosen not to use the term *stylometry*, which is the common term for quantitative authorship attribution: most of the textual variables used in *stylometry* are not wholly stylistic.

2.12 Word Position[51]

In addition to the many investigators who have examined word frequencies, there have also been a number of investigators that have chosen to count words only when they occur in a particular position in a text. This branch of quantitative authorship attribution is known as *positional stylometry* and is the brainchild of the prolific investigator of authorship, the Reverend Andrew Q. Morton.

Morton first began to publicize his methods and results in a series of lectures delivered across the United Kingdom and North America in 1963.[52] In these lectures, Morton claimed to have discovered empirical evidence that Saint Paul had written only *Romans, I and II Corinthians, Galatians*, and *Philemon*—nine fewer Epistles than are traditionally thought to have been written by the Apostle to the Gentiles. Morton's research made headlines: a front page article in the November 7th 1963 edition of the *New York Times* announced his discovery. His original arguments have since been republished many times, including in his 1965 paper for the Royal Statistical Society, which was referred to earlier in this review.[53] At this time, Morton's methods involved using the Chi-square statistic to compare the values of three basic types of textual measurements. The first two types of measurements, based on sentence-length and function word frequency, were not unique at the time and have already been considered in this review (see section 2.4 and 2.11). But Morton was also analyzing the frequency of function words in particular positions in relationship to the front or back of a text's sentences. In this case, and in many others, Morton examined the frequency of the Greek conjunction *de* ("but") as the second or third word of a sentence. It was this type of positional measurement that would be at the core of Morton's research for the next thirty years.[54]

While Morton's work was influential and received much attention from the public, it almost immediately came under heavy criticism. Herdan's scathing 1965 attack has already been referenced, but perhaps the most damning critique came from the

[51] The main text is Morton (1978). For good reviews see Oakman (1980), Holmes (1998), Love (2002).

[52] For example, see Morton (1965c), which is based on a 1963 lecture at University of Saskatchewan.

[53] Where Morton removed *Philemon* from the list. See also Morton (1965b); Morton & James McLeman (1966); Levison & Morton & Wake (1966); Morton & McLeman (1964); Morton & Levison (1966).

[54] See also O'Donnell (1966) for the frequency sentences that begin with conjunctions as an indicator of authorship. See also Kenny (1986), Neumann (1990) and Ledger (1995), for additional quantitative attribution studies of the Paulines.

Reverend John Ellison (1965), who demonstrated that when Morton's methods were applied to James Joyce's *Ulysses* or to Morton's own works, they determined that each of these texts were the product of multiple authors. Despite these critiques, Morton continued his research and moved on to new attribution problems and gained new associates along the way. For example, in his 1968 publication with Levison and Winspear, Morton used his techniques to argue that Plato did not write the *Seventh Letter*.[55] And Morton, Levison, Winspear and Michaelson returned to Ancient Greek texts in their 1971 book *It's Greek to the Computer*, where they applied their techniques to a number of classical cases of disputed authorship, including the *Iliad,* the *Odyssey,* the *Seventh Letter,* the *Axiochus*, the *Epinomis,* the *Timaeus Locrus* and the *Nicomachean Ethics*. But the watershed year was 1972: Morton and Sidney Michaelson—a professor of computer science at the University of Edinburgh—published no less than five different articles, many of which introduced new techniques to the field, such as the ratio of genitive to non-genitive forms of the pronoun *autos* (1972a), the size of the interval between successive occurrences of the conjunction *kai* (1972b), and the frequency of words occurring at the end of sentences (1972c).[56] The last two techniques are indicative of the type of positional measurements that Morton and Michaelson were beginning to develop further. A useful introduction and defense of Michaelson and Morton's battery of positional attribution techniques for the Ancient Greek language are presented in their 1973 paper "Positional Stylometry."[57]

In the mid-seventies, Morton and his collaborators shifted from Greek to English texts. In order to attribute English texts Morton had to introduce a new type textual measurement, for, as Ellison had demonstrated, Morton's sentence-position tests could not attribute texts written in an uninflected language like English, with its necessarily strict word order. Morton's new method was based on the frequency of sequences of words or what are known as collocations. For example, the previous sentence ends with the four-word collocations *are known as collocations*.[58] Morton began by applying his

[55] For criticism of this argument see Brandwood (1969).

[56] For critique of Morton's last word method see P.F. Johnson (1974).

[57] Here Morton and Michaelson identify two basic types of words in regards to their range of potential sentence-positions: an *isotropic* word occurs throughout a sentence, while an anisotropic words, such as *de*, occurs mainly in a limited range of positions, such as at the beginning of a sentence.

[58] This method is described in Michaelson, Morton and Hamilton-Smith (1979). See also Morton 1978.

new technique to classic cases of disputed English authorship: for example, the London *Observer* published an article in 1976 entitled "Word Detective Proves the Bard wasn't Bacon." These studies brought Morton much public attention and soon he was being asked by lawyers to testify on behalf of their clients.[59] In his most celebrated case, Morton appeared at Old Bailey as an expert for the defense, and testified that the defendant's confession was written by multiple authors and was thus forged by the police. The court agreed and the charges were dropped and Morton was famous.

In 1978, Morton published *Literary Detection*, one of the most important texts in the history of quantitative authorship attribution. In this book, Morton's presented his most refined set of techniques, which were based on the frequency of function words in particular sentence positions, function word collocations, and proportional pairs. In addition to defending and describing his techniques, he also applied his methods to a series of problems, including the authorship of *Pericles*, which Morton argued was written by Shakespeare in its entirety, and the second half of Jane Austen's unfinished novel *Sanditon,* which had been completed by an admirer after Austen's death. Morton also demonstrated that his methods could distinguish between the works of Sir Arthur Conan Doyle and his imitators, who continue to add to the canon of Sherlock Holmes.

The publication of *Literary Detection* inspired many attribution studies. One of the most ardent supporters of positional stylometry was Thomas Merriam, whose interest in grapheme frequencies, word ratios and neural networks has already been noted in this history, but who was also an early practitioner and defender of Morton's techniques. In a series of studies (1979, 1980, 1982), Merriam applied Morton's methods to the uncertain edges of the Shakespearean Canon, including *Henry VIII, Sir Thomas More,* and *Pericles.*[60] In another notable study, the economists O'Brien and Darnell (1982) used Morton's collocation method—focusing on frequent collocations whose second word is either *the* or *be*—to investigate a number of important cases of disputed economic writings, including the works of Sir Josiah Child, the 17th century East India Company tycoon.[61]

[59] See Totty & Hardcastle & Pearson (1987) for a review of Morton's methods as applied to statements.
[60] See also Merriam (1989), in which he uses Morton's Proportional pair method on the *Federalist Papers.*
[61] See also Elliott and Valenza (1996), where word position variables are used in addition to other tests.

While Morton's new techniques were being widely applied, they were also being widely criticized, especially by M.W.A. Smith. At first, Smith had subscribed to Morton's theories,[62] but he soon became their main detractor as he discovered flaws in both Morton's techniques and his experimental design.[63] Smith's major assault came in 1985, in a pair of papers published in respected journals, where he pointed out defects in Morton's techniques which he claimed led to unreliable results. In the second paper, Smith focused on Merriam's application of Morton's techniques in the three papers cited above.[64] Smith followed up these attacks with an article (1987a) in which he criticized a new positional technique introduced by Morton (1986), which was based on the frequency of *hapax legomena* at the beginning or ending of sentences and in collocations with certain high frequency function words. But even after Smith became a critic of Morton's research, he was still one of the major contributors to field of positional stylometry. For example, in a 1988 article, Smith introduced a new test of authorship based on the first words of a play's speeches. However, while his tests were similar to Morton's, Smith's results were not: after testing his method on a corpus of Elizabethan and Jacobean playwrights, Smith used the technique to argue that Wilkins and not Shakespeare wrote the first two acts of *Pericles.*[65]

In the early 1990s, Morton shifted his focus again and, along with Sidney Michaelson, began to rely on cusum charts—graphs that plot how the mean value of a textual measurement changes over the course of a text—to attribute authorship. Morton had mentioned these charts in passing at least as early as 1966 (see Morton and McLeman 1966:96), and had even produced examples of these charts in his 1968 paper with Levison & Winspear, as evidence that Plato did not write the *Seventh Letter.* As Morton *et al.* wrote (1968:320)

> The graphs here appended known as cumulative summary plots, or more familiarly, *cusum* plots, are of very great interest....These graphs are, so far we are aware, the first application of the method to literature.

[62] See for example Smith (1983) where he suggests using contingency tables larger than 2x2.

[63] See for example Smith (1984, 1994).

[64] Merriam then defends the studies in Merriam (1986, 1987, 1997), and then Smith attacks again (1987c).

[65] See Smith (1987b, 1989a) for other studies of *Pericles*, and see Smith (1987d) for an investigation of *The Revenger's Tragedy.*

At that time, Morton and his associates provided almost no explanation of how these graphs were generated and did not even label the graphs that they reproduced.[66] But the method was developed further and placed at the forefront of Morton's approach to authorship attribution with the publication of two papers: Morton & Michaelson (1990), and Morton (1991).[67] The method was now being used to graph two cusum charts that measured the changing values of related textual measurements, on the assumption that they should run parallel to one another, as long as the author of the text remained the same. For example, in Morton and Michaelson (1990), one cusum chart graphed the changing average sentence-length over the sentences of a text, while a second cusum chart graphed the changing values of the number of words in each sentence which begin with a vowel. Assuming that this second measurement is stable across the works of an author, whenever the average sentence-length rises or falls, then the value of the second measurement should follow. In order to use this method to examine a text whose consistency is in doubt, the two cusum graphs are generated and then superimposed: if the two graphs are roughly parallel, then the text is deemed to be the work of a single author; but if the two graphs diverge, then a new author is posited to have taken over at that point of divergence. In principle, the method can also be used to test if two or more texts were written by the same author by conjoining the texts and seeing if the cusum charts diverge at the juncture.

While the method does sometimes produce impressive results, and was soon being used by defense lawyers to analyze their client's incriminating confessions (see Morgan 1991, Campbell 1992), its accuracy was disputed by a number of investigators. The detractors questioned the theory behind the technique—as David Holmes writes "the underlying assumption regarding the consistency of habits within the utterances of one person is false" (1998:114)[68]—and the lack of any clear definition of what exactly

[66] For more early use of cusum see Bee (1971, 1972), Morton, Levison, Winspear and Michaelson (1971), Michaelson & Morton (1972e), Michaelson & Morton & Wake (1978), and Morton (1978), where he points to how the cusum charts that show the change in the occurrence of particle *de* at the beginning of sentence in the Fourth Gospel, as evidence that the story about the women taken in adultery is anomalous (1978:88).

[67] See also Farringdon *et al.* (1996), for an introduction to the cusum technique; and see Merriam (2000) for a recent study of *Edward III* which uses cusum charts to analyze word frequencies and metrical variables.

[68] For other critics of cusum see: Canter (1992), de Haan and Schils (1993), Sandord et al (1994), Holmes & Tweedie (1995), Hardcastle (1993).

constitutes an adequately parallel set of graphs. But the technique and Morton's most embarrassing failure came in 1993 when he was challenged on live British television to attribute texts that he had never seen. The result was disastrous: despite his impressive statistics and his fancy computer graphics, Morton could not distinguish between the writings of a convicted felon and the Chief Justice of England (Sams 1994:472).

Today Morton's methods are often dismissed: for example, in his short but informative review of quantitative authorship attribution, Holmes writes "if stylometry had its 'dark age' then surely this must be it" (1998:113). But Morton's techniques have started to be reintroduced. Most notably, there have been a number of recent studies based on collocation frequency, although now the measurement often goes under different names: Hoover (2002) looks at "frequent word sequences"[69] and Clement & Sharp (2003) call it only a "naive Bayes classifier." Unfortunately, these investigators rarely give credit where credit is due, preferring perhaps not to have their work associated with the research of a largely discredited investigator of authorship. No matter the success of these methods, in my mind this is far from fair. The Reverend Andrew Morton deserves to be recognized as one of the most prolific, influential and innovative investigators in the history of quantitative authorship attribution.

2.13 N-Grams

A recent addition to our stock of textual measurements are character-level *n-grams—n-*length sequences of contiguous characters. For example, if we consider only the twenty-six letters of the English alphabet and the space, then in any English text there are 27 possible 1-gram-types (*A, B,..., Z,_),* 729 possible 2-gram-types *(AA, AB,...*), 19683 3-gram-types, etc.[70] It is very easy to split a text into n-grams: e.g. the phrase *author unknown*, which consists of fourteen characters, contains fourteen 1-gram tokens (*A, U, T, H, O, R,_, U, N, K, N, O, W, N), thirteen 2-gram tokens (AU, UT, TH, ..., WN), twelve 3-gram tokens (AUT, UTH,...,R_U, ..., OWN*), etc. In general, a text that contains x number of total characters will contain a total of $x - (n + 1)$ n-gram tokens.

[69] See also Hoover (2003), where he looks at what he calls "collocations" which includes both the standard sequence of (two) words, but in this case also includes pairs words that occur near each other.

[70] Often other characters, such as punctuation marks, digits, symbols, tabs and paragraphs are included into the analysis.

Using the frequencies of n-grams as an indicator of authorship seems to have first been proposed by William Ralph Bennett in his 1976 text *Scientific and Engineering Problem-Solving with the Computer.*[71] In this study, Bennett used two-gram frequencies to attribute the works of a number of authors. This early study would have been forgotten if not for the scholarship of Keselj, Fuchun, Cercone and Thomas (2003), who referenced this text in their own n-gram-based attribution study. Here, Keselj *et al.* emphasized that one of the main advantages to an n-gram-based attribution method is that it can be applied, in principle, to texts written in any language. Keselj *et al.* tested their method using a range of n-gram sizes (one- to ten-grams) and achieved success distinguishing between sets of English, Greek and Chinese authors.[72] In particular, they found that the seven English authors were distinguished perfectly when the frequency of the 1500 to 2000 most common 6-grams or 7-grams were compared. In a second 2003 study from the same group of researchers, Peng, Schuurmans, Keselj and Wang used a larger range n-gram profiles to attribute a similar collection of datasets and achieved similar results.

While an impressive degree of accuracy was achieved in all of these studies, it must be acknowledged that their results are based on questionable experimental design. Specifically, the sets of possible authors that these researchers considered (e.g. Keselj *et al.* 2003: Bronte, Burroughs, Carroll, Cleland, Dickens, Haggard, Irving, Shakespeare) span such a wide range of dialects, registers, eras and subjects that it is impossible to predict if their method would be capable of distinguishing between a more stylistically and thematically homogeneous set of possible authors—the type of set that one would expect to find in an actual case of disputed authorship. For example, to demonstrate that a measurement can discriminate between the works of Shakespeare and Dickens is to demonstrate nothing, except perhaps that the measurement can distinguish between Elizabethan plays and nineteenth century novels, which is a very different (and, presumably, easier) classification task. One must therefore assume that the results of these studies are inflated.[73] In fact, when these investigators tested their methods on the

[71] N-grams are even more commonly used in meaning- and language-based text classification, see Cavner & Trenkle (1994) for examples of both.

[72] They used the set of possible authors compiled and investigated by Stamatatos, Fakotakis & Kokkinakis (2001).

[73] See section 4.2.3, where I discuss proper experimental design in greater detail.

more strictly controlled corpus of Greek authors, which are all contemporary journalists, their results were far less impressive.

However, in a 2003 study, Clement and Sharp conducted a more careful and extensive evaluation of n-gram-based attribution methods. These researchers compared the performance of one- through twenty-five-gram-based attribution methods on a set of possible authors composed of online reviews of the same movies. But despite their careful experimental design, Clement and Sharp's results were similar to those of the researchers referenced above: they achieved their best results when analyzing the frequency of eight-grams.

In a related type of authorship study, Khmelev and Tweedie (2001) experimented with grapheme two-grams by constructing Markov models where each possible sequence of two letters is represented by the probability that the second letter follows the first in a particular author's writing sample. They tested this method on a selection of questionable datasets including the *Federalist Papers* (which is a poor dataset as it is a case of disputed authorship), the small mystery writer corpus used in Baayen *et al.* (1996) (see section 2.14), and a corpus consisting of 387 texts written by 45 stylistically heterogeneous possible authors. Overall, the method appears to be fairly successful, achieving 74.4% accuracy when distinguishing between the set of 45 authors; but, in all cases, these datasets are too uncertain or too small or too variable to allow any strong general conclusions to be drawn from this study.[74]

Despite the often disappointing experimental design of these studies, n-gram frequencies are likely to be good general indicators of authorship because they are sensitive to the frequencies of a wide range of linguistic units that are often examined in attribution studies, such as graphemes, punctuation marks, words and collocations. N-grams are also sensitive to an author's use of affixes. For example, the frequency of the verbal past tense suffix would affect the frequency of the 3-gram *ed_*, and an author's use of the negative prefix would affect the frequency of the 3-gram *_un*. It would thus seem that the main advantage of using an n-gram-based attribution method is that it is sensitive

[74] See also Kukushkina, Polikarpov & Khmelev (2002), who examine sequences of graphemes and grammatical classes (see section 2.14 for more on this study); Kjell (1994), who used 2-grams and neural nets to examine the *Federalist*; and Forsyth & Holmes (1996), who use a Chi-square based analysis of both the most frequent 2-grams and two other methods ("Progressive Pairwise Chunking" and "Monte-Carlo Feature Finding") that basically examine frequently occurring n-grams of any length.

to many different types of textual patterns. But more careful evaluations of these techniques are needed to verify this assumption.

2.14 SYNTAX

There are many potential indicators of authorship that can only be accessed through a syntactic analysis of a text. For example, our language offers stylistic optionality in word order (*the man walked quickly* vs. *the man quickly walked* vs. *quickly the man walked*, and *on the hill he is fast* vs. *he is fast on the hill*) that has a minimal effect on measurements such as n-gram frequency and collocation frequency, and that has no effect on measurements such as word frequency, vocabulary richness, grapheme frequency and sentence-length. There are also many syntactic measurements that are likely to be good thematic indicators of authorship, such as the relative frequency of adjectives: some authors, no matter what subject they are writing about, tend to provide more information about nouns than other authors do.[75] Despite the promise of such syntactic measurements, there have only been a few attribution studies that have directly analyzed the syntax of a text.

One of the earliest quantitative attribution studies to analyze syntax was O'Donnell's 1966 investigation of Stephen Crane's unfinished final novel *The O'Ruddy* (see section 2.5). In this study, O'Donnell considered the values of a number of syntactic measurements including the relative frequency of verbs, adjectives, adverbs, clauses, dependent clauses, simple sentences, past-participle sentences, and impersonal constructions. O'Donnell chose to examine these syntactic features because (1966:109)

> These variables directly relate to "how it is said," the essence of style. Ideas expressed in a series of independent clauses rather than in a series of simple sentences, or subordinated by a participle phrase rather than by a dependent clause, have a different impact on the reader even though the content is identical.

O'Donnell then used discriminant analysis—in what appears to be the first use of advanced multivariate techniques in attribution studies—to compare the values of these syntactic measurements, in addition to measurements of punctuation mark frequency and

[75] Adjective frequency is not usually a stylistic property of an utterance. For example, the differences between the noun phrases *a dog*, *a black dog* and *a big black dog* are purely thematic: the differences is in the amount of information that has been conveyed about the dog.

average word- and sentence-length, in the works of Crane and Barr. Based on this evidence, O'Donnell concluded that Barr had not written as much of the novel as he had claimed.

A similar syntactic approach to authorship attribution was taken by the linguist Jan Svartvik, in *The Evans Statement* (1968), which is perhaps the most shocking case of disputed authorship in the quantitative attribution canon. This case involves a series of contradictory confessions that were allegedly dictated to the police by John Evans, who was subsequently executed for the brutal murder of his pregnant wife and infant daughter. Svartvik focused on two of the statements in particular. In the first statement, Evans recounts how his pregnant wife had died when their neighbor John Christie had botched an abortion, and how he had then given up his child for adoption. But in the second statement, Evans purportedly admits that he had murdered his wife and daughter in his home. The second statement stood and Evans hung, but four years later the police returned to the home that the Evans and Christie families had shared, and found the body of Christie's wife. Soon Christie was charged with his wife's own murder and was executed as well, but before he died he admitted that he had strangled Evans' wife and that Evans was innocent of the crime. Svartvik set out to compare the style of the two contradictory Evans statements to see if he could demonstrate that the second statement had been altered by the police in order to convict Evans of a crime that he did not commit.

Svartvik's analysis of the Evans statements began with a qualitative reading of the texts. Svartvik noticed stylistic discrepancies in the two statements: he found the first statement to be much more colloquial, containing features such as double negatives and non-standard verb conjugation. Svartvik also noticed internal variation in the damning second statement: the beginning and the end of the second statement, which agree in content with the beginning and the end of the first statement, were also written in a similar colloquial style; but the middle of the statement, where Evans allegedly admits to the crime, is far more formal. To confirm his qualitative impressions of the texts, Svartvik looked at the proportions of six types of clauses: free clauses, clauses with mobile relator, clauses with immobile relator, clauses with elliptic subject linkage, conjunctional clauses, and relative clauses. Over the three sections of the second statement, Svartvik found a great deal of variation: e.g. 20.5% of the clauses in the middle section had mobile relators, while only 3.9% and 4.2% of the clauses in the outside sections of the confession were of this type. Furthermore, when Svartvik split the first statement into three sections, the proportion of clause types was relatively stable and far more similar to outside sections of the second statement. Based on this type of syntactic evidence, Svartvik concluded that Evans was probably innocent of the crime.

In a less gruesome study, the classicist Anthony Kenny (1986) used ninety-nine syntactic measurements to help resolve some of the cases of disputed authorship found within the *New Testament*: Did the author of the *Gospel of Luke* also write *Acts*? Did the author of *John* also write *Revelations*? And, of course, which of the Pauline Epistles were actually written by Saint Paul? Kenny's ninety-nine measurements were mostly based on a fine-grained part-of-speech typology. For example, not only did Kenny consider the frequencies of verbs, but he also looked at the frequencies of verbs in the active, passive, middle, or deponent voice; and in the present, future, imperfect, aorist, perfect, or pluperfect tense; and in the subjunctive, optative, imperative, and infinitive moods. He then used the correlation coefficient to compare the values of these sets of syntactic measurements in the relevant books of the *New Testament.* Based on this technique Kenny concluded that the author of the *Gospel of John* probably did not write *Revelations* and that Luke may very well have written *Acts.* And while Kenny found variation across the Paulines, in this case he hedged his conclusion, writing that (1986:100)

> What is to be said of the authorship of the Epistles is in the end a matter for the Scripture scholar, not the stylometrist. But on the basis of the evidence in this chapter for my part I see no reason to reject the hypothesis that twelve of the Pauline Epistles are the work of a single, unusually versatile author.

This is a result, of course, that differs significantly from the conclusions of Morton and W. B. Smith.[76]

In an even more complex study of part-of-speech frequency, Kukushkina, Polikarpov & Khmelev (2002) used Markov chains to model an author's preferences for sequences of word classes: i.e. given one word class how likely is it that a second word class will follow? Kukushkina *et al.* tested this method using 18 word classes to attribute a large and fairly well-designed corpus of 385 texts written by 82 Russian authors.

[76] See also Foster (1989), for another example of parts-of-speech frequencies as indicators of authorship.

Overall, Kukushkina *et al.* found that their technique attributed 235 of these texts correctly (61%). One other very interesting result of this study also deserves to be mentioned: Kukushkina *et al.* found that texts that were translated into Russian could also be attributed to their original author, even if they had been translated by different interpreters.[77]

In another recent attribution study, Stamatatos, Fakotakis and Kokkinakis (2001) used the output of a natural language processing tool to classify texts based on authorship. In particular, Stamatatos *et al.* used a "syntactic chunker" that they had developed (see Stamatatos, Fakotakis and Kokkinakis 2000) to determine the values of twenty-two textual measurements, including the relative frequency and average lengths of noun phrases and verb phrases, which were compared using a discriminant analysis. They then tested their method on a carefully constructed corpus of Modern Greek journalists. In order to judge the usefulness of their technique, they set a baseline using discriminant analysis to classify the texts based on an analysis of the frequency of the 50 most common Greek function words, achieving a 74% success rate when distinguishing between the texts of ten possible authors. When their syntax-based method was tested on the same dataset, they achieved a better success rate of 81%.

Finally, in an example of what can be accomplished with a fully parsed corpus, Baayen, van Halteren and Tweedie (1996) examined the potential of using syntactic rewrite rule frequency as an indicator of authorship. In order to obtain a sufficiently large sample of syntactically annotated texts, Baayen *et al.* made use of the fully-parsed Nijmegen corpus. A register controlled test corpus was then compiled by extracting an equal number of parsed samples from two 1960 English crime novels written by two different authors. They then tested two rewrite-rule frequency methods to see if they could discriminate between the texts of these two authors. The first technique used a principal components analysis to compare the frequencies of the fifty most frequent rewrite-rules in the test corpus, and the second technique used a principal components analysis to compare the values of five different vocabulary richness measures—*K, D, R,*

[77] See also Radday (1970), for a more limited analysis of word class transition frequencies as an indicator of authorship, where only four different word classes were recognized—nouns, verbs, others, and stops—for a total of sixteen possible word class pairs; and see Koppel & Schler (2003) for a study of email authorship that looks at part-of-speech 2-grams as an indicator of authorship.

S, W—which were calculated for the texts' rewrite-rule frequency distributions. These two methods were based on two popular methods discussed earlier in this history—the multivariate function word frequency technique used by Burrows (1988) and the multivariate vocabulary richness technique used by Holmes (1992)—which were also used to set a baseline against which to compare the performance of their new rewrite rule techniques. While in all four cases the techniques were able to distinguish between the author's works, Baayen *et al.* concluded that (1996:129)

> We interpret this result as confirming our initial intuition 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 constructions leads to a higher discriminatory resolution.

Hopefully, as automated parsing technology continues to improve, similar studies will be conducted with larger sets of possible authors.

2.15 Summary

In this chapter, I have presented a history of quantitative authorship attribution. Now that the various types of textual measurements have been identified, in Chapter 3, I will select a large set of the most important of these indicators, and explicitly describe how their values can be calculated and compared in order to attribute an anonymous text. These measurements will then be evaluated in Chapters 4 and 5.

3 ATTRIBUTION ALGORITHMS

3.1 INTRODUCTION

The purpose of this project is to compare the most commonly used sets of textual measurements in quantitative attribution in order to determine which are the best general indicators of authorship. A fair comparison requires that each set of measurements be tested in the same manner and on the same dataset. In this chapter, I describe the basic quantitative attribution algorithm and the sets of textual measurements that will take their turn at its core.[78] The dataset and the testing procedure will be introduced in Chapter 4.

3.2 INPUT

All quantitative methods for authorship attribution are similar: they are procedures for selecting a text's most likely author from a set of possible authors by comparing the values of one or more textual measurements in that text to their corresponding values in each possible author's writing sample. All attribution algorithms therefore take as input an anonymous text and a set of possible author writing samples. Generally, each possible author is represented by multiple writing samples so as to increase the likelihood that any patterns found in those texts are characteristic of that author.

The goal of an attribution algorithm is to determine which possible author's writing sample is the most similar to the anonymous text, in terms of the values of a particular set of textual measurements. It is the responsibility of the investigator to provide the algorithm with a valid set of possible authors, which includes the anonymous

[78] I have implemented and tested all thirty-nine of the algorithms described here (and their variants, for a total of ninety-three different sets of textual measurements) using the programming language PERL.

text's actual author. It is also the responsibility of the investigator to test the attribution algorithm before it is applied, to ensure that that particular set of textual measurements can distinguish between that particular set of possible authors. One might envision a more powerful attribution algorithm capable of determining if a particular author wrote an anonymous text based only on an analysis of the anonymous text and that author's writing sample, but this is not possible: no matter how similar that possible author's writings samples are to the anonymous text, some other possible author's texts may be an even better match. Unless the attribution algorithm is given a chance to compare the anonymous text to *all* its possible authors' writing samples, there is no way to know if a more similar possible author exists. If the investigator is not confident that he has identified a valid set of possible authors, then he must interpret his results with care. Authorship attribution is not analogous to DNA testing and it never will be: different people cannot have identical DNA, but different people can produce identical utterances.

3.3 TEXTUAL PREPARATION

In this study, some textual preparation is undertaken before the input texts are compared. Most significantly, all quoted passages are deleted from the input texts. This step is taken because all the texts analyzed in this study are newspaper articles, where any quoted passage is unlikely to have been uttered by the columnist. Fortunately, it is fairly easy to automate the removal of quotations from a text if its author, like all the columnists in this study, uses double quotation marks as their primary markers of quotation.[79]

The following procedure is used to de-quote a text. First, if the same symbol marks both the opening and the closing of a quotation (i.e. "), then quotation marks need to be disambiguated. This is accomplished by replacing an ambiguous quotation mark with an open quotation mark (i.e. “), if it comes after a space or an open bracket or at the beginning of a paragraph; and by replacing any other ambiguous quotation mark with a close quotation mark (i.e. ”), except when it comes after a dash (i.e. —"), in which case it is replaced with a close quotation mark only if this sequence is followed by a space. The text is then de-quoted by reading through each paragraph one character at a time, copying

[79] It is much more difficult to delete quoted passages that are delimited by single quotation marks because the single primary quotation mark and the apostrophe can be ambiguous.

each character, until an open quotation mark is encountered, at which point the copying of the text is paused. Copying is resumed only after the next primary close quotation mark is read. The trick is to distinguish between primary and embedded quotation marks, because if there are additional sets of double quotation marks within a set of double quotation marks, then the first close quotation mark that follows the first open quotation mark does not mark the end of that quotation. In order to distinguish primary and embedded quotation marks, once the first open quotation mark is read and the copying of the paragraph is paused, any open quotation marks that follow are counted and the copying of the paragraph is not resumed until an equal number of closed quotation marks have been read. Finally, if the quoted passage contains a sentence boundary (see section 3.4.3 below for a definition of a sentence boundary), then the quoted passage is replaced by a period, in order to preserve the text's sentence structure. Once the end of the paragraph is reached, any open quotation marks are closed before moving on to the next paragraph. The text must be de-quoted one paragraph at a time because in a series of quoted paragraphs each paragraph must begin with an open quotation mark, but only the last paragraph must end in a close quotation mark.[80]

The only time that this procedure for text de-quotation fails is in the case of punctuation errors by the author, incomplete texts, and quotation marks that are being referred to explicitly by the author. It is assumed that each of these cases is sufficiently rare to be ignored.

A few additional steps of simple textual preparation were also undertaken. First, any datelines or bylines were removed from the input texts. Second, spacing was cleaned up so that paragraphs are separated by only one newline and words are separated by only one space, and all tabs were also removed from the text.

[80] Where a paragraph is defined as a string of characters that occurs between two consecutive newlines, the start of the text and the first newline, or the last newline and the end of the text.

3.4 TEXTUAL MEASUREMENTS

3.4.1 INTRODUCTION

In order to determine which possible author's writing sample most resembles the anonymous text, it is necessary to extract some information from each of these texts to compare. In quantitative authorship attribution, this information consists of the values of a set of one or more textual measurements, where a textual measurement is defined as a function of the frequencies of one or more strings of characters in a text. It is necessary that these textual measurements be defined in an unambiguous manner so that their values can be mechanically and consistently calculated for any text. In this section, I define the sets of textual measurements that will be tested in this paper. All of these measurements are based on the nine most popular indicators of authorship: word-length, sentence-length, vocabulary richness, grapheme frequency, punctuation mark frequency, word frequency, word position frequency, collocation frequency and n-gram frequency.[81]

3.4.2 WORD-LENGTH

Two word-length-based sets of textual measurements are tested here. The first measurement is a text's average word-length in graphemes, which is calculated by dividing the total number of grapheme-tokens in the text, by the total number of word-tokens in the text. The second set of measurements is a text's word-length profile, which consists of the relative frequency of one-letter words, two-letter words, three-letter words, etc. in the text. The relative frequency of each word-length-type is calculated by dividing the total number of word-tokens of that length in a text, by the total number of word-tokens in the text. Various forms of the word-length profile are tested here, which differ in terms of the number of word-length-types included in the profile. But, in all cases, only those word-length-types that occur in at least two of every possible author's writing samples are included in the profiles. This restriction is necessary because of the statistical technique (see section 3.5) and the testing procedure (see section 4.3) being used in this study. This same restriction is imposed on all the textual profiles tested here.

[81] The complete list of textual measurements and all their variants are listed in the Appendix. The Appendix also includes the specific members of each set of textual measurements, which depend on the corpus of possible authors introduced in Chapter 4.

In order to count the number of words in a text, it is necessary to define a word. The approach taken here is to define a word as a continuous string of graphemes and/or digits. Word boundaries consist of all other characters including punctuation marks, symbols and spaces. Therefore two strings of graphemes and/or digits connected by a hyphen (-) or a contraction (') are considered to be two separate words (e.g. *does, not, doesn, t*). Furthermore, by this definition, two identical strings of graphemes and/or digits with different meanings or functions are considered to be tokens of the same word-type (e.g. *table* as a noun or verb, or *to* as a preposition or an infinitival marker, or *Jack* as a proper noun and *jack* as a common noun).

3.4.3 Sentence-Length

Four sets of sentence-length-based textual measurements are tested here. The first measurement is a text's average sentence-length in words, which is calculated by dividing the total number of word-tokens in the text, by the total number of sentences in the text. The second set of measurements is a text's word-based sentence-length profile, which consists, for example, of the relative frequency of one- to five-word sentences, six- to ten-word sentences, etc. in the text. The relative frequency of each sentence-length-type is calculated by dividing the total number of sentences in a text that fall in that range of sentence-lengths, by the total number of sentences in the text.[82] Various forms of the word-based sentence-length profile are tested here, which differ in terms of the range and number of sentence-length-types being counted. The third measurement is a text's average sentence-length in characters, which is calculated by dividing the total number of characters in the text, by the total number of sentences in the text. The fourth set of measurements is a text's character-based sentence-length profile, which consists, for example, of the relative frequency of one- to thirty-character sentences, thirty-one- to sixty-character sentences, etc. in the text. The relative frequency of each sentence-length-type is calculated by dividing the total number of sentences in a text that fall in that range of sentence-lengths, by the total number of sentences in the text. Various forms of the

[82] Sentence-length-types are defined as an interval of sentence-lengths in order to avoid sentence-length-types with a frequency of zero. This allows for larger sentence-length profiles.

character-based sentence-length profile are tested here, which differ in terms of the range and number of sentence-length-types being counted.

In order to count the number of sentences in a text, it is necessary to define a sentence. The approach taken here is to define a sentence as a string of characters that occurs between two sentence boundaries, where a sentence boundary consists of a question mark, an exclamation mark or a sentence-final period followed by whitespace and a capital letter.[83] But before a text can be split into sentences, it is necessary to delete all those periods that are followed by a space and a capital letter, but which mark a proper name and not the seam of two sentences. For the most part, such periods are fairly easy to identify and delete. First, if any instance of the abbreviations *mr., mrs., ms., dr., rev., r., ed., gen., lt., rt.* and *hon.* is discovered in the text, its period is deleted because the following word is almost always a proper name. Second, the final period in the abbreviations *vs., v., e.g.* and *i.e.* are deleted because, while it is rare for a capital letter (i.e. a proper name) to follow these abbreviations, it is even rarer for these abbreviations to occur at the end of a sentence. Third, any period is deleted that is preceded by a capital letter and then a space (i.e. *_X._*), because if the next character is a capital letter, then it is probably the start of a surname, and the first capital letter is probably an initial.[84]

This procedure for sentence-splitting will fail in at least two cases: it will not split a text where it should whenever initials or acronyms occur at the end of a sentence but no second period follows (e.g. *They work at N.A.S.A..* vs. *N.A.S.A.*), and it will split a text where it should not whenever a proper name follows other abbreviations which do commonly occur at the end of sentence (e.g. *Tim gave Kim et al. Jim's application*). Overall these and other exceptional cases are assumed to be infrequent enough that they will not have much effect on any of the sentence-length-based measurements.[85]

3.4.4 Vocabulary Richness

Eleven Vocabulary Richness measurements are tested here. Where N is the total number of word-tokens in the text, V is total number of word-types in a text, V_i is the total number

[83] Periods that precede quotation marks can be ignored at this point because they have already been deleted during de-quotation.

[84] This is essentially the sentence-splitting algorithm presented in Manning & Schütze (1999).

[85] This procedure would fail far more often if the corpus was not composed of carefully written monologues. Furthermore, when splitting sentences, it does not matter if the sentence is grammatical

of word-types that occur exactly *i*-times in a text, and p_v is the relative frequency of the *v*-th word-type in a text.

(1)	*TTR.*	=	V / N
(2)	K	=	$10^4 (\sum i^2 V_i - N) / N^2$
(3)	R	=	$V / \sqrt{N}$
(4)	C	=	$\log V / \log N$
(5)	H	=	$(100 \log N) / (1 - V_1/V)$
(6)	S	=	V_2 / V
(7)	k	=	$\log V / \log (\log N)$
(8)	LN	=	$(1 - V^2) / (V^2 \log N)$
(9)	ENT	=	$-100 \sum p_v \log p_v$
(10)	W	=	$N^{V \wedge - a}$

In addition, various forms of *W* are calculated for different values of *a*. The eleventh vocabulary richness technique tested here is a limited Type-Token Ratio, which is calculated in the same way as the regular Type-Token Ratio, except that it is based on only the first *n*-number of words in every text, where *n* is the length of the shortest writing sample in the corpus of possible authors. This measurement is made because the Type-Token Ratio is known to be very sensitive to text-length.

One can group these measurements into two basic types: *TTR, R, C, k, LN,* and *W* are based only on the number of word-types and word-tokens in a text; whereas *K, H* and *S* are based on the grouped word frequency distribution, which depends on the number of word-types that occur once, twice, thrice, etc. in a text.

3.4.5 GRAPHEMES

Four sets of grapheme-frequency-based textual measurements are tested here, where a grapheme is defined as one of the twenty-six letters of the English alphabet (no distinction between capital and lower case letters is made). The first set of measurements is a text's grapheme profile, which consists of the relative frequencies of the twenty-six English graphemes in the text. The relative frequency of each grapheme-type is calculated by dividing the total number of tokens of that grapheme-type in a text, by the total number of grapheme-tokens in the text. The second set of measurements is a text's word-position grapheme profile, which consists of the relative frequencies of graphemes occurring in a particular position in relationship to the front or the back of the text's words. The relative frequency of each grapheme-position-type is calculated by dividing the total number of word-tokens with that grapheme in that position, by the total number of word-tokens in the text. Various forms of the word-position grapheme profile are tested here, which differ in terms of which word-position is being analyzed, including the first grapheme in a word, the second grapheme in a word, and the last grapheme in a word. The third set of measurements is a text's multi-word-position grapheme profile, which consists of the relative frequencies of graphemes occurring in various positions in relationship to the front and/or the back of a text's words. The relative frequency of each grapheme-position-type is calculated as above. Various forms of the multi-word-position grapheme profile are tested here, which differ in terms of which word-positions are being analyzed. These multi-word-position profiles are combinations of the single-word-position grapheme profiles. For example, a profile based on the frequency of graphemes in the first three positions of a word would contain three separate measurements of the frequency of the grapheme *A* (i.e. its frequency in the first, second and third position of a text's words). The fourth set of measurements is a text's word-internal grapheme profile, which consists of the relative frequencies of graphemes occurring anywhere within the text's words. The relative frequency of each grapheme-type is calculated by dividing the total number of word-tokens in the text that contain that grapheme, by the total number of word-tokens in the text.

3.4.6 Words

One set of non-positional word-frequency-based textual measurements is tested here. This set of measurements is a text's word profile, which consists of the relative frequency of a set of words in the text. The relative frequency of each word-type in the profile is calculated by dividing the total number of tokens of that word-type in a text, by the total number of word-tokens in the text. Various forms of the word profile are tested here, which differ in terms of the number of words included in the profile. Because of the frequency restrictions imposed by the statistical and experimental design of this study, the largest word profile tested here contains all those words that occur in at least two of every possible author's writing samples, and the smallest word profile tested here contains all those words that occur in every possible author's writing samples. Other word profiles tested here fall somewhere in between these two extremes.

As the frequency restriction is raised not only does the number of words in the profile decrease, but the proportion of content to function words also falls. This is because the most frequent words in English are function words. By increasing the frequency restrictions, it is thus possible to only analyze the frequency of function words. This is significant, because it is generally assumed that function words are better indicators of authorship than content words. This theory will be tested here by comparing the results of algorithms based on different sized word profiles.

The standard linguistic distinction between content (or lexical) and functional (or non-lexical) words is that function words are members of closed word classes, whereas content words are members of open word classes (Schachter 1985, Biber *et al* 1999), where a word class is defined as a set of words that share a similar grammatical distribution over a set of sentences (Fries 1952). Open word classes are distinguished from closed word classes based on the rate at which new words are added to the class: a closed word class admits very few new members over time, whereas the size of an open class like nouns "is in principle unlimited, varying from time to time and between one speaker and another" (Robins 1964:230). For example, consider the difference between an open class like nouns and a closed class like pronouns. Nouns are added to our language everyday: some refer to newly discovered or invented things, while others add another synonym or near-synonym to our lexicon, replete with new stylistic or thematic connotations. The openness of the class is reflected in the size of the class and the way that we, as individual speakers of a language, learn new nouns as we gain new knowledge. On the other hand, there are fewer than fifty pronouns in English, and I do not believe that I have learnt a new pronoun, with a new meaning, since I was a child. That is not to say that our inventory of pronouns is eternal. Language changes. The pronoun *thy* rarely rolls off the modern English tongue, while the compound pronouns *he or she* and *he/she* have been recently added to the class in order to offer more politically correct variants of the gender-neutral *he.*[86] Nonetheless, the membership of a closed word

[86] Indeed, there have been a few very unsuccessful attempts to introduce new gender-neutral pronouns, such as *tey* (Miller & Swift 1971). This is further evidence of the closeness of the pronoun word class.

class like pronouns is significantly more stable than the membership of an open word class like nouns.

In the English Language, around eighteen different word classes (and major subclasses) are recognized. A standard word class taxonomy is presented in Table 1, where the closed word classes are marked in bold type.[87]

TABLE 1 ENGLISH WORD CLASSES

WORD CLASS	**EXAMPLE**
Nouns	Home, rock, dizziness...
Pronouns	I, me, you, they, himself...
Adjectives	Big, smart, other, red...
Determiners	The, his, that, all, what...
Verbs	Sleep, jump, wish, change...
Modal Verbs	Will, can, may, should...
Auxiliary Verbs	Be, is, are, was, had, do...
Adverbs	Quickly, easily, happily...
Adverbial Particles	Too, also, then, just...
Prepositions	In, of, to, by, for, at...
Conjunctions	And, or, if, but, that ...
Numerals	One, two, three, four...
Ordinals	First, second, last...
Negative Marker	Not, ‘t
Existential Marker	There
Infinitival Marker	To
Genitive Marker	‘s
Interjections	Um, ah, like, eh ...

3.4.7 PUNCTUATION

Five sets of punctuation-frequency-based textual measurements are tested here. The first three sets of measurements are variations of a text's punctuation mark profile, where the relative frequency of eight punctuation-mark-types (. , : ; - ? (‘) are calculated by dividing the total number of tokens of that punctuation-mark-type in the text, by the total number of word-tokens, or the total number of character-tokens, or the total number of punctuation-mark-tokens in the text. The fourth and fifth sets of measurements consist of a text's punctuation and grapheme profile and a text's punctuation and word profile, where the relative frequency of each punctuation-mark-type is calculated by dividing the total number of tokens of that punctuation-mark-type in the text, by the total number of

[87] This taxonomy is based on the system used to annotate the British National corpus (Leech *et al.* 2001). A full explanation can be found at ***http://www.comp.lancs.ac.uk/ucrel/bnc2/bnc2guide.htm***

graphemes or words in the text; and the relative frequency of each grapheme- and word-type is calculated as described in the two sections immediately before this one. In all five of these sets of measurements, all punctuation marks are assumed to have occurred at least once in every text, otherwise many punctuation marks would have to be excluded from the analysis because some authors abstain from using them altogether.

3.4.8 WORD POSITIONS

Two sets of word-position-based textual measurements are tested here. The first set of measurements is a text's sentence-position word frequency profile, which consists of the relative frequencies of words occurring in a particular position in relationship to the front or the back of the text's sentences. The relative frequency of each word-position-type is calculated by dividing the total number of sentence-tokens in a text with that word in that position, by the total number of sentence-tokens in the text. Various forms of the sentence-position word profile are tested here, which differ in terms of which sentence-position is being analyzed. The second set of measurements is a text's multi-sentence-position word profile, which consists of the relative frequencies of words occurring in various positions in relationship to the front and/or the back of a text's sentences. The relative frequency of each word-position-type is calculated as above. Various forms of the multi-sentence-position word profile are tested here, which differ in terms of which sentence-positions are being analyzed. These multi-sentence-position word profiles are combinations of the single-sentence-position word profiles.

3.4.9 COLLOCATIONS

Two sets of collocation-based textual measurements are tested here, where a collocation is defined as a sequence of two or more word-types. The first set of measurements is a text's two-word-collocation profile, which consists of the relative frequencies of two-word-collocations in the text. The relative frequency of each collocation-type is calculated by dividing the total number of tokens of that collocation-type in a text, by the total number of two-word-collocation-tokens in the text, which is equal to the total number of word-tokens in the text minus one. The second set of measurements is a text's three-word-collocation profile, which consists of the relative frequencies of three-word-collocations in the text. The relative frequency of each collocation-type is calculated by dividing the total number of tokens of that collocation-type in a text, by the total number of three-word-collocation-tokens in the text, which is equal to the total number of word-tokens in the text minus two. Larger collocations are too infrequent to be analyzed.

3.4.10 N-Grams

Eight sets of character-level n-gram-based textual measurements are tested here, where an n-gram is defined as a sequence of two or more characters (including graphemes, digits, spaces, newlines and punctuation marks). The first set of measurements is a text's two-gram profile, which consists of the relative frequency of two-grams in the text. The relative frequency of each two-gram-type is calculated by dividing the total number of tokens of that two-gram-type in a text, by the total number of two-gram-tokens in the text, which is equal to the total number of character-tokens in the text minus one. The remaining seven sets of measurements consist of a text's three- through nine-gram profiles, where the relative frequency of each n-gram-type is calculated by dividing the total number of tokens of that n-gram-type in the text, by the total number of n-gram-tokens in the text, which is equal to the total number of character-tokens in the text minus (n-1). Various forms of each of these n-gram profiles are tested here, which differ in terms of the number of n-grams being analyzed. As was the case with word profiles, this is accomplished by varying the minimum number of each author's writing sample in which each n-gram must occur at least once to be included in the profile.

3.5 Comparison & Output

As outlined above, the basic attribution algorithm takes as input an anonymous text and a set of possible author writing samples. After preparing the input texts, the algorithm reduces each to a textual profile, which consists of a set of textual measurements plus their specific values in that text. The algorithm then compares the anonymous text's profile to each possible author's profile. But first, because each author is represented by multiple writing samples, their profiles are combined to form one profile for each possible author. This is accomplished by averaging the values of each textual measurement across each of that author's profiles. The anonymous text's profile is then compared to each possible author's profile to determine which pair is the closest match.

To compare two profiles it is not sufficient to simply sum the absolute differences of the values of each corresponding pair of textual measurements. Rather, it is necessary to somehow scale the results so that the value of a few measurements does not overwhelm all the others. For example, if profiles containing the relative frequency of the fifty most common words were compared in this manner, only the frequency of the four or five most frequent words (*the, and, of, a, to*) would matter.

In this study, I have therefore chosen to use the Chi-square statistic (χ^2) to compare the textual profiles (Cochran 1952, 1954, Siegel 1956, Woods & Fletcher 1986, Oakes 1998). The Chi-square test is a simple statistic used to measure the goodness-of-fit between the observed and expected frequencies of a set of independent nominal categories. In general, the Chi-square test is used to determine if a sample, represented by a set of observed frequencies, could have been drawn from a particular population, represented by a corresponding set of expected frequencies. In particular, given *m*-number of nominal categories, the goodness-of-fit between a set of observed frequencies ($O_1, O_2, \ldots, O_m$) and a set of expected frequencies ($E_1, E_2, \ldots, E_m$) is calculated using the following formula:

$$(12) \qquad \chi^2 \quad = \quad \Sigma((O_i - E_i)^2 / E_i) \qquad i = 1, 2, 3, \ldots, m$$

The Chi-square test is thus computed by subtracting each category's expected frequency from its observed frequency, taking the square of this value, dividing the result by the category's expected value, and finally by summing the values of this operation for each of the categories being compared.

The lower the Chi-square value, the more confident we may be that the sample could have been drawn from that particular population: if the two sets are identical, then the Chi-square value is zero. To interpret a non-zero Chi-square value, a critical Chi-square table is usually consulted: depending on the number of categories and the chosen level of significance, the table specifies whether the Chi-square value is low enough to conclude that the differences between the two sets are insignificant (i.e. only a result of sample error), or if the differences are too large for the sample to be considered to have been drawn from that population.

In attribution studies, the Chi-square test is often used to compare the observed frequencies of a set of textual measurements in an anonymous text to the sets of frequencies that would be expected if the text were written by a particular possible author, based on an analysis of that author's writing samples (e.g. Brinegar 1963, Morton 1965a, Kenny 1978, O'Brien & Darnell 1982, Usher & Najock 1982, Forsyth & Holmes 1996, Chaski 2001). In essence, the Chi-square test allows the investigator to determine from which possible author's population of texts the anonymous text was most likely drawn. In this study, the Chi-square test is used specifically to compare the anonymous text's profile to each possible author's profile. The algorithm then outputs the list of possible authors ranked by ascending Chi-square value, where the author associated with the smallest Chi-square value is deemed to be the anonymous text's best match.

Three points should be made about this application of the Chi-square test. First, the critical Chi-square table is not consulted. Instead the algorithm simply outputs a ranking of the possible authors. If it were not assumed at the outset that the set of possible authors contained the text's true author, then it would be necessary to interpret the Chi-square values in order to determine if the anonymous text was similar enough to be considered the work of the best-matched author. But there is no need to consult the Chi-square table here because it is assumed that the anonymous text's actual author is one of the possible authors whose writing samples are being analyzed: assuming that the set of possible authors is valid and that the algorithm was properly tested, the possible author with the smallest Chi-square value is the algorithm's selection.

Second, while the Chi-square test is meant to compare frequencies, most of the profiles being compared in this study consist of relative frequencies. These relative frequencies could easily be transformed into frequencies by being multiplied by the length of the anonymous text, but it is perfectly reasonable to use the Chi-square test to compare the relative frequencies directly because the rankings of possible authors will be the same in either case. This is demonstrated in Table 2, where hypothetical relative frequency profiles are compared directly in the first part of the table, frequencies per thousand words in the second part of the table, and frequencies per four million words in the third part of the table.

TABLE 2 HYPOTHETICAL CHI-SQUARE CONTINGENCY TABLES

		N = 1			N = 1,000			N = 4,000,000	
Word	**Text**	**Aut 1**	**Aut 2**	**Text**	**Aut 1**	**Aut 2**	**Text**	**Aut 1**	**Aut 2**
The	0.056	0.051	0.055	56	51	55	224000	204000	220000
And	0.026	0.021	0.031	26	21	31	104000	84000	124000
To	0.021	0.025	0.027	21	25	27	84000	100000	108000
Of	0.028	0.028	0.023	28	28	23	112000	112000	92000
A	0.022	0.024	0.026	22	24	26	88000	96000	104000
χ^2		0.0025	0.0039		2.4873	3.8603		9949.4	15441
C		0.0498	0.062		0.0498	0.062		0.0498	0.062

Clearly, in all cases the resultant ranking of possible author is identical. Furthermore, in all cases the Chi-square value of Author 2 is 1.55 times larger than the Chi-square value of Author 1. This consistency is reflected by the fact that, no matter the value of N, the authors are associated with the same coefficient of contingency (C), which is calculated using the following formula.

$$(13) \qquad C \quad = \quad \sqrt{}(\chi^2 / (N + \chi^2))$$

On the other hand, the Chi-square values are not preserved, but this is not a problem because the Chi-square table is never consulted.

Third, no restriction is placed on the values that the textual measurements can assume, except that only measurements whose value is over zero in every possible author's profile may be compared, in order to eliminate any divisions by zero (although zero values are acceptable in the anonymous text).[88] Crucially, the standard restriction, that only measurements whose expected frequencies are over five may be included in the analysis, is ignored. Usually this restriction is necessary because the Chi-square table would otherwise be unreliable, but in this case the Chi-square table is never consulted. This is significant because it allows for a larger number of textual measurements to be included in the profiles. For example, notice that whereas all the measurements used in the example above would be valid for N = 1000 (as all have an expected frequency of over 5), none of the measurements would be valid for N = 1, but yet in both cases the exact same ranking is produced when the values of all five textual measurements are compared. Conforming to this restriction would therefore needlessly limit the number of measurements that could be analyzed. This fact has not been acknowledged in the past

[88] This restriction taken together with the testing procedure introduced in the next chapter has the result of only allowing textual variables in the profile that occur in at lest two of each authors writing samples.

and so the number of potential measurements that have been analyzed using the Chi-square statistic has always been kept unnecessarily small.

Finally, it should be noted that not only is the Chi-square test used in this study to compare profiles containing measures of relative frequency, but the Chi-square test is also used to compare those univariate profiles that only contain the value of a single measurement (i.e. average word- and sentence-length and vocabulary richness). A more appropriate and straightforward method for comparing two of these univariate profiles would be to simply calculate their absolute difference: there is only one measurement being compared and therefore there is no need to scale the comparison. In fact, this method for comparing univariate profiles was tested, but it was found to produce results that were nearly identical (±0.4%) to the results obtained using the Chi-square test. For the sake of consistency, the Chi-square test was therefore used, in the manner outlined above, to compare all the profiles.[89]

3.6 SUMMARY

In this chapter, I have introduced thirty-nine fully-defined attribution algorithms, which differ only in terms of the types of textual measurements whose values they compare. By evaluating these attribution algorithms and their variants, I will therefore be able to evaluate the sets of textual measurements upon which they are based. The results of these tests will be presented in Chapter 5, where I will also present the results of testing algorithms based on combinations of the individual attribution algorithms described here.

[89] Frankly, the few univariate methods tested here will prove to be the least accurate of all the indicators of authorship anyway, and so it would appear to matter very little how their values are compared.

4 EXPERIMENTAL DESIGN

4.1 INTRODUCTION

To evaluate the authorship attribution algorithms introduced in Chapter 3, and the sets of textual measurements at their core, it is necessary that each be tested in the same manner and on the same dataset. In this chapter, I describe the experimental design of this study: I explain how the *corpus of possible authors* was compiled and how this dataset was used to test the performance of these thirty-nine types of attribution algorithms. The results of this experiment are presented in Chapter 5.

4.2 THE CORPUS OF POSSIBLE AUTHORS

4.2.1 CORPUS COMPILATION

In empirical linguistics, a *corpus* is defined as "a finite collection of machine-readable text sampled to be maximally representative of a language or variety" (McEnery & Wilson 1996:177). A *representative* corpus, one "that approximates the closest to the population from which it is drawn" (*ibid*:178), is necessary so that observations made of the corpus will hold true of the variety of language that the corpus represents. Because it will be argued below that a rigorous test of an attribution algorithm's performance requires that the algorithm be tested on a highly representative corpus of possible authors, it is important that the compilation of the corpus of possible authors be informed by basic corpus linguistic methodology.

I here propose three basic steps to compile a representative corpus. First, the variety of language that the corpus represents must be unambiguously defined. Second, this variety of language must be divided into sub-varieties. And third, texts must be gathered that exemplify each of these sub-varieties. Before this process can be described in greater detail, it is first necessary to define *language*, *a variety of language* and linguistic *style*—three terms that are needed for this discussion. William Dwight Whitney offers a definition of *language* that is particularly suitable for the empirical linguist (1901:6):

> All the accessible forms of human speech, in their infinite variety, whether still living in the minds and mouths of men, or preserved only in written documents, or carved on the scantier but more imperishable records of brass and stone.

Language is thus the unimaginably large corpus that contains every utterance mankind has ever produced. This universal corpus is characterized by those universal grammatical properties that all utterances share (e.g. the presence of morphemes and nouns). A *variety of language* is a smaller corpus that is composed of all the utterances that share some additional non-universal grammatical properties and that are the product of some extra-linguistic situation.[90] Both of these conditions are necessary: not all grammatically-defined collections of utterances are valid varieties (e.g. the collection of all utterances containing a prime number of morphemes is not a valid variety because these utterances are not unified by any common situation) and not all situationally-defined collections of utterances are valid varieties (e.g. the collection of all utterances produced in the rain is not a valid variety because these utterances are not unified by any non-universal grammatical properties). A valid variety of language is rather a corpus of utterances that is defined by the pairing of a particular grammar with a particular extra-linguistic situation. Finally, the *style* of an utterance is the set of grammatical properties that allow it to be recognized as a product of a particular extra-linguistic situation or, equivalently, as a member of a particular variety of language (Crystal & Davy 1969).

The first step to compiling a representative corpus is therefore to define a linguistically significant extra-linguistic situation. This is a complex task: there are many situational variables that affect the style of an utterance and many others that do not. In order to determine if a particular situational variable is linguistically significant, one must

[90] Where situation is defined as "the setting in which a use of language takes place" (Crystal 1987: 430).

compare utterances produced in situations that differ only in the value of that single situational variable. If these utterances can be distinguished based on the values of one or more *linguistic variables*—i.e. linguistic units (e.g. phonemes, morphemes, syntactic structures) with at least two variant but semantically equivalent forms (Chambers & Trudgill 1980:60)—then that situational variable is linguistically significant. To define a variety of language one must specify the values of the linguistically significant situational variables of the situation in which its utterances are produced.

Fortunately, sociolinguists have identified and organized the most significant situational variables into more general *dimensions of linguistic variation*, which can be used to define a valid variety of language. While there is no definitive typology of situational categories, there is a fair deal of agreement across the systems that have been proposed. For example, in their 1969 study of English style, David Crystal and Derek Davy provide a catalogue of seven major "dimensions of situational constraints": Dialect (social background of interlocutors: e.g. Australian English, Black English Vernacular), Time (e.g. Late Middle English, Modern English), Medium (e.g. speech, writing), Participation (direction of communication: e.g. monologue, dialogue), Province (occupational setting: e.g. legalese, sports announcing), Status (relationship of interlocutors: e.g. formal, informal), and Modality (the "purpose" of communication: e.g. poetry, prose, lecture, note, essay).[91] I believe that this is the most insightful of these systems, and the typology that I propose below is essentially a simplification of this system. However, the most influential typology was introduced by Dell Hymes (e.g. 1974:54-62), who identified sixteen situational dimensions: Message Form, Message Content, Setting (time and place of the situation), Scene (cultural significance of the situation), Speaker, Addressor, Hearer/Audience, Addressee, Purposes/Outcomes, Purposes/Goals, Key (manner and tone of the situation: e.g. mock vs. serious, perfunctory vs. painstaking), Channel (Medium), Forms of Speech, Norms of Interaction, Norms of Interpretation, and Genres.[92] Another influential and more parsimonious system was

[91] Crystal & Davy also propose two more dimensions: *Individuality* (e.g. sound of voice, handwriting), which seems out of place as this is not associated with any linguistic patterns; and *Singularity*, which is their default category for those significant situation types with fall into none of their other categories.

[92] Some of these categories seem superfluous and others out of place, at least given Crystal's definition of *situation,* which has been adopted here. For example, Message Form and Message Content are aspects of the utterance and not of the situation in which an utterance is produced.

proposed by M. A. K. Halliday, which consists of three basic situational dimensions (1978: 62): Field (the purpose of the act communication), Tenor (the participants in the act communication) and Mode (Medium, Genre, Key).[93]

Without necessarily making a theoretical claim, in my opinion any such system can be reduced with little loss of precision to a more manageable and complete set of three basic and distinct situational dimensions of linguistic variation: Dialect, Time and Register. In this study, the variety of language that a corpus represents is therefore defined by specifying the values of these three determinants of linguistic style.

The dialect of a variety of language is defined in terms of the *social* situation in which its utterances are produced (Crystal & Davy 1969, Crystal 1987, Chambers & Trudgill 1980, Petyt 1980, Ferguson 1994). In any act of communication there are two major social variables that affect the style of the utterances: the speaker and the audience. The significance of the speaker is obvious: if one compares utterances produced in identical situations except that they are produced by different speakers, then the utterances would differ in the values of multiple linguistic variables. But similarly, if one compared utterances produced in identical situations except that they were produced for different audiences, then the utterances would differ grammatically as well. In particular, one would find that the style of the utterances would tend to shift towards the style of the audience. For example, many speakers from the lower-classes affect an upper-class speaking style when being interviewed for a white-collar job. To define the social situation in which an utterance is produced one must therefore specify its speaker and its audience.

It is a more complex task to define the dialect of a larger variety of language that is produced by multiple speakers for multiple audiences. In such cases, one cannot select any set of speaker-hearers and call the set of utterances that they produce a dialect. Instead, one must select a set of speaker-hearers that produces a grammatically homogeneous set of utterances. Such a set of speaker-hearers is known as a *speech community*, which is a concept best understood by conducting Leonard Bloomfield's classic thought experiment (1933:46):

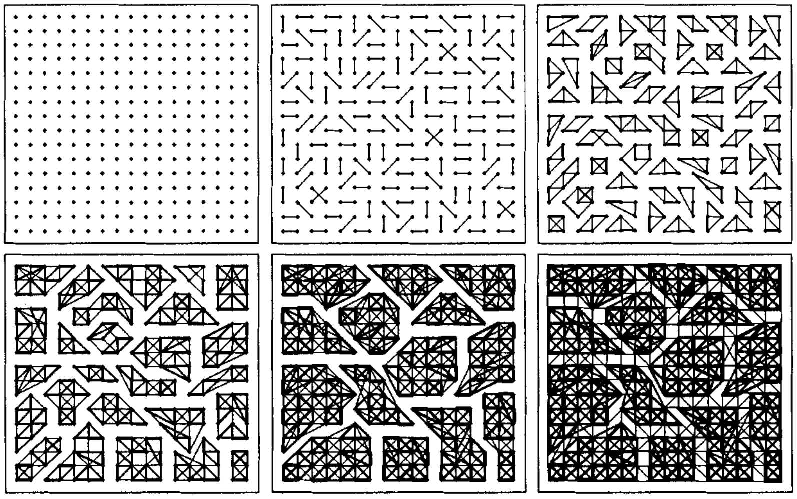
[93] See Biber (1995) for a discussion of various descriptive frameworks of extra-linguistic situations.

> Imagine a huge chart with a dot for every speaker…and imagine that every time any speaker uttered a sentence, an arrow were drawn into the chart pointing from his dot to the dot representing each one of his hearers.

Now imagine how the lines of communication would amass over time: the nodes would not be connected at random. Rather, the speakers would cluster together to form *speech communities*—clusters of speakers who communicate more often with each other than with the rest of language's speakers.

In Figure 1, I present a hypothetical series of Bloomfield's charts, where the thickness of a link represents the activity of that connection.

FIGURE 1 A HYPOTHETICAL SPEECH COMMUNITY



A speech community appears in these charts as a network of speaker-hearers. This network may range in size from a single link connecting two speaker-hearers to the complex network that links every speaker-hearer in the chart. All but the smallest of speech communities are therefore composed of smaller speech communities—even more active networks of communication that are embedded inside the larger speech community. When one defines the dialect of a variety of language, one may therefore specify almost any sized set of speaker-hearers, as long as one specifies a set of speaker-hearers who communicate more often with each other than with the rest of language's speakers. But the smaller the speech community that one defines, the more specific the grammar that characterizes its utterances will be. This is because the utterances of a smaller speech community will always conform to the grammar of any larger speech community in which it is embedded, while also conforming to those additional grammatical properties that only characterize its own utterances.

In order to define a valid speech community and hence the dialect of a variety of language, it is easiest to specify the social background of its speaker-hearers. This is because speech communities are composed of speaker-hearers with similar geographic (e.g. neighborhood, city, region, nation) and demographic (e.g. family, class, age, race, ethnicity, education level, religious) backgrounds: if one were to examine an annotated form of Bloomfield's chart where social information was provided about each speaker, it would be clear that speakers with similar social backgrounds tend to communicate more often with each other than without. This social specification may be as general or as precise as one likes, but the more general the specification, the larger and less active the speech community, and hence the less grammatically homogeneous that community's utterances will be.

The reason that utterances produced by a speech community exhibit a greater degree of grammatical homogeneity than the universal set of utterances is because, as Ferguson writes (1994:18)

> A group that operates regularly in a society as a functional element...will tend to develop identifying markers of language structure and language use, different from the language of other social groups.

This tendency toward grammatical homogeneity can be explained in two ways. First, social groups take advantage of the inherent variability of language to distinguish themselves from other social groups. Second, while there is great motivation for us all to speak the same variety of language, so that we may communicate with as many people as possible, the inherent variability of language and the inherent limitations of the human mind only allow us to speak a similar variety of language as the people with whom we communicate most often.

While the dialect of a variety of language is determined by the social situation in which its utterances are produced, the era of a variety of language is determined by the *temporal* situation in which its utterances are produced. It is necessary to specify the span of time over which a variety of language was produced because the form that an utterance takes depends in part on the point in time in which it was uttered. This is the basic observation of historical linguistics (Lehmann 1962, Labov 1994). As Winfred P.

Lehmann writes: “In historical linguistics we study differences in languages between two points in time” (1962:3). The length of a linguistic era can range anywhere from the time it takes to produce a single utterance to the entire history of the human tongue, but, once again, the more narrowly that this dimension is defined, the more grammatically homogeneous the set of utterances produced in that situation will be.

Finally, the register of a variety of language is determined by the *communicative* situation in which its utterances are produced (Biber 1995, Ferguson 1994, Martin 2001, Eggins & Martin 1997). At the most basic level, register is determined by the medium over which its utterances are transmitted—“the intervening substance through which impressions are conveyed to the senses” (New Oxford English Dictionary). The register dimension can thus be divided immediately into two major categories: spoken and written registers.[94] These two registers may be divided further by identifying more specific media-types: spoken media include the face-to-face conversation, the telephone conversation, and the voicemail message; and written media include the letter, the email and the text message. Unlike the arbitrary grammatical properties that characterize dialects and eras, the grammatical properties that characterize registers are a result of the properties of the register, which cause certain linguistic variants to be favored over others. For example, utterances produced in the text message medium will tend to contain more abbreviations and creative spellings, and fewer redundant function words and inflections, than utterances produced in the email medium, because text messages are harder to type.

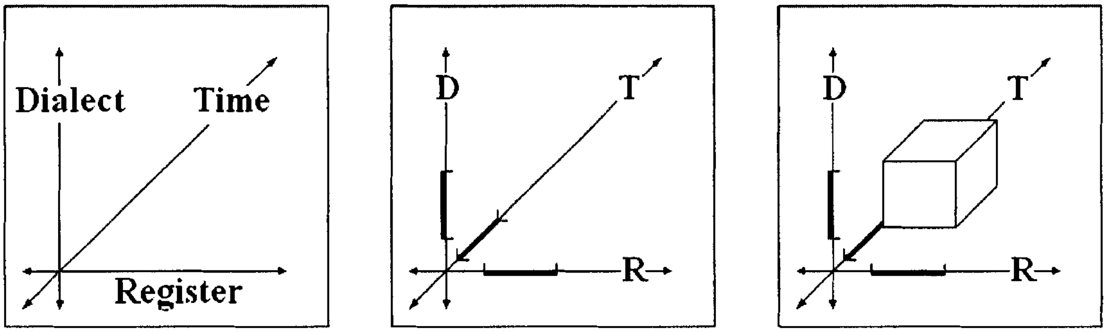
But the communicative situation in which an utterance is produced does not depend solely on its physical properties, it also depends on its functional properties. This is because the function for which an utterance is produced can cause certain linguistic variables to be preferred over others. For example, consider the linguistic and functional differences between the poetry and prose registers: imagine if one were to compare utterances of poetry and of prose that were produced in otherwise identical situations (i.e. by the same writer, for the same audience, at the same point in time, and in the same medium). If these utterances can be distinguished it is because of differences in the values of some set of linguistic variables. For instance, the two sets of utterances might

[94] Though language can be communicated by other means: e.g. in English we also use signing and rebuses.

differ in the rate of contraction, especially if the poems were composed in strict metrical verse, where contractions are often necessary if the poet's lines are to conform to the meter of the poem. The function of the poem—in this case, to have a pleasant, regular and recognizable rhythm—thus causes certain linguistic variants to be favored over others. When one defines the register of a variety of language one must therefore specify both the physical and functional characteristics of the communicative situation in which its utterances are produced.

By specifying a dialect, era and register one can define a linguistically significant situation and thus a valid variety of language. The fact that these three dimensions of linguistic variation seem to be sufficient to define a variety of English is convenient as it allows us to envision this process in terms of our familiar three spatial dimensions, as illustrated in Figure 2.

FIGURE 2 TO DEFINE A VARIETY OF LANGUAGE



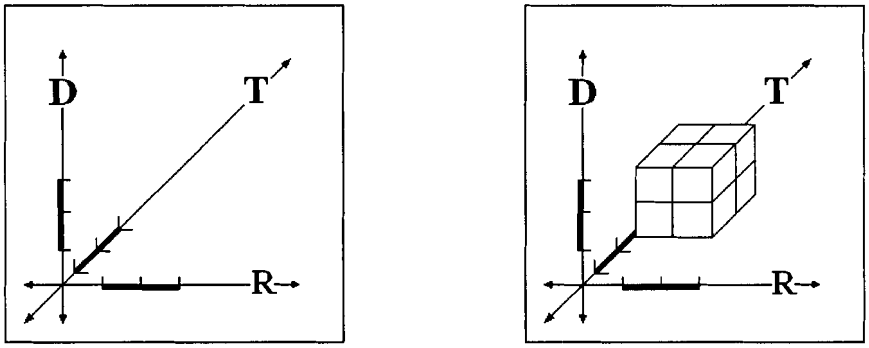
For example, valid varieties of language include Modern English (dialect: the entire English dialect continuum[95]; time: approximately 1500 A.D. to present; register: all), Spoken 1970s Lower Class Black Philadelphian English (dialect: lower class Black Philadelphian Speech Community; time: 1970-79; register: informal spoken language), and John Kerry's Presidential Concession Speech (dialect: John Kerry addressing his supporters and the broadcast audience; time: Wednesday, November 4, 2004; register: live televised presidential concession speech).

[95] In my opinion, the best technical definition of *a language* is as the variety of language spoken by an entire dialect continuum (e.g. English, Arabic). The fact that traditional language such as German and French are reduced by this definition to dialects of the Germanic and Romance languages is perhaps unfortunate, but the term cannot be rigorously defined in any other way (see Grieve 2004).

Once the variety of language that a corpus represents has been defined, it must then be divided into sub-varieties so that each sub-variety may be represented in the corpus. This is an important step: a corpus's degree of representation depends on how finely and how accurately it has been divided into sub-varieties. For example, one would not want to make generalizations about the grammar of the entire English language based on a corpus that is composed of the grocery lists of a man from London. Instead, one would want to base a grammatical theory of the English language on a corpus that samples from as many varieties of English as possible.

In order to divide any variety of language into sub-varieties, it is simplest to divide its already defined dialect, era and register into sub-dialects, sub-eras and sub-registers. This process is illustrated in Figure 3, where each dimension has been divided in two to produce eight sub-varieties.

FIGURE 3 TO DEFINE THE SUB-VARIETIES OF A VARIETY OF LANGUAGE



For example, if one is creating a rather modest corpus of 20th Century Standard British English Prose, one should sample texts from at least the first and second halves of the century, the northern and southern standard dialects, and the fiction and non-fiction registers. In total this creates eight distinct sub-varieties of 20th Century Standard British English Prose: early northern fiction, early northern non-fiction, early southern fiction, early southern non-fiction, late northern fiction, late northern non-fiction, late southern fiction, and late southern non-fiction. Of course, a finer division of 20th Century Standard British English Prose would be preferable, so that we may be more confident that any grammatical properties found in that corpus are characteristic of that variety of language.

In general, the empirical linguist should attempt to divide the variety as finely as possible, and should defer to a more finely divided corpus if one becomes available.

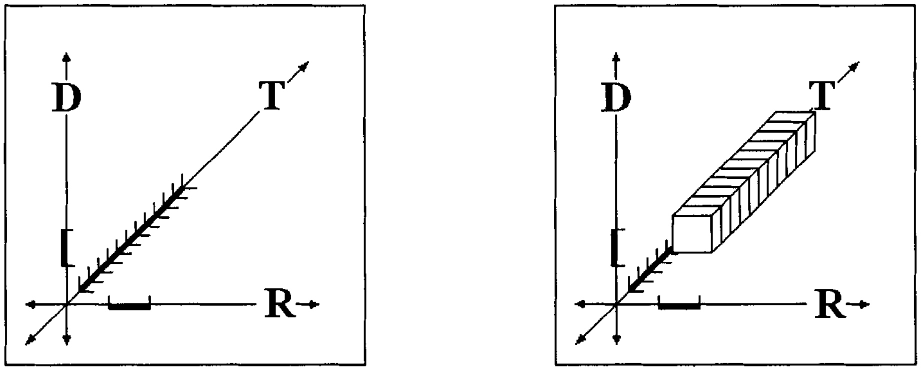
Once the variety of language has been divided into sub-varieties, the corpus may finally be compiled by gathering texts that exemplify each of the sub-varieties, i.e. texts that are produced in the particular dialect, era and register that define that sub-variety. Any variety of language is a population of utterances and so when one compiles a corpus one is taking a sample of that population. Thus, in general, the more texts that are gathered, the more representative the corpus will be. The number of utterances that represents each sub-variety should also be proportional to the actual number of utterances in that sub-variety. This is because many of the types of grammatical properties that characterize varieties of language are probabilistic: e.g. Americans don't always say *y'all* and Canadians don't always say *you guys*, but they are each more likely to use these particular variants of the second person plural pronoun than the other. If the number of texts that represents each sub-variety is not proportional to its size, then the corpus may not be truly representative.

4.2.2 AUTHOR-BASED CORPUS COMPILATION

Now that the basic principles of corpus construction have been established, they may be used to direct the compilation of the author-based corpora that will be used in this study. These corpora, each of which represents a variety of language in which a possible author writes, will provide both the writing samples and the "anonymous" texts upon which the attribution algorithms will be tested. If the performance of these algorithms is to be accurately gauged, then these corpora must be compiled with consistency and care.

It is not a trivial matter to define the variety of language in which an author writes. Most writers interact with multiple readers, at multiple times, in multiple media, and with multiple purposes, and so one must decide which of an author's many varieties the author-based corpus will represent. When attributing an anonymous text, it is both unnecessary and unsound to compile an author-based corpus that represents the variety of language that encompasses all that author's written utterances: the anonymous text is the product of a single situation and so each author-based corpus should be composed of texts produced in the same register, for the same audience, and around the same point in time as the anonymous text. Otherwise, the investigator might get false negatives: when the anonymous text is compared to the corpus of its actual author they may not match because of stylistic variation that is the product of differences in audience or register or time. This said, the variety of language that an author-based corpus represents should not be too narrowly defined: an author-based corpus should always contain multiple writing samples so as to increase the likelihood that any patterns found in that corpus are characteristic of that author. Optimally, the time dimension alone is extended, because even if one writing sample were sufficient to attribute an anonymous text, the time dimension would have to be extended, or else the text's actual author could never be included in the analysis, as he would have produced no text at that point in time, other than the text that is being attributed. The composition of the ideal author-based corpus is illustrated in Figure 4.

Figure 4 The Ideal Author-Based Corpus



But it is not always possible to compile such a narrowly-defined author-based corpus. This is because the investigator does not always know the situation in which the anonymous text was produced, and because the possible author may not have produced any texts in that particular situation—except, perhaps, the text that is being attributed. In such cases, the investigator must endeavor to compile author-based corpora that are as similar as possible to the anonymous text, by extending the value of the three dimensions of linguistic variation far enough to allow sufficient number of texts to be sampled.

In this attribution study, where the goal is to compare the performance of multiple attribution algorithms, and where there is thus no anonymous text to direct the compilation of the author-based corpora, these corpora must still be defined in the narrowest of terms if they are to provide a realistic test of the attribution algorithm's performance.

Clearly, the dialect of each of the author-based corpora will be defined in terms of a social situation consisting of a single author; but to define the narrowest of dialects one must also specify a stable audience. The indivisible dialect, produced by a single speaker for a single audience, is known as an *idiolect*. This term was introduced by Bernard Bloch, and while it is often used by linguists to refer to the variety of language that encompasses the totality of an individual's utterances (e.g. see Hockett 1958), this is not how Bloch intended the term to be used. Rather, he defines an idiolect as (1948:7)

> The totality of the possible utterances of one speaker at one time in using a language to interact with one other speaker.

In this study, each author-based corpus represents an idiolect. This was accomplished by only selecting authors who write for the London *Telegraph*, and by only sampling these authors' regular *Telegraph* opinion columns. Admittedly, the readership of a newspaper column is never entirely stable, but because the readership is fairly stable and because the readership is so large and so anonymous that the columnist never knows its exact composition, the columnist will usually treat his readership as a stable audience, and will thus usually write in a stable dialect. This is not to say that a columnist is never aware of and thus affected by a shift in his audience: the newspaper may be in the process of trying to appeal to a new segment of the population, or the columnist may have recently been provided with the result of a survey which has changed his impression of his readership, or the columnist may know that a friend who does not usually read the newspaper will be reading it today. Such social factors could cause a columnist's dialect to change, but usually a columnist's audience is stable. This is especially true of established columnists who write regular columns for major newspapers with a large and dedicated audience, such as all the *Telegraph* columnists used in this study.

By choosing to compile author-based corpora that represent the variety of language in which *Telegraph* opinion columnists write, the register of these corpora has also been defined in the narrowest of terms: the *Telegraph* opinion newspaper column is a very specific type of register. This register is particularly well-suited for attribution studies because newspaper columns are plentiful and in the public domain, and because the *Telegraph* offers a large online archive from which machine-readable texts can be downloaded for free. The newspaper opinion column register also provides texts that are carefully written, of a relatively consistent length, and that are short enough (usually five hundred to two thousand words) to provide a challenging test for the algorithms.

While the dialect and register dimensions of the author-based corpora used in this study have now been defined in the narrowest of terms, if these corpora are to contain more than a single text, then their time dimensions must be extended. Fortunately, newspaper columnists are some of the most regular and prolific writers of published English texts, and so in this study the time span could remain relatively short, while still allowing for a large number of texts to be included in each author-based corpus. In particular, I have chosen to include forty columns in each author-based corpus, which for most of the columnists requires that texts be sampled from a one-year time span. Usually this time span ranges from January 2004 to January 2005, but in all cases the columns were written between the years 2000 and 2005.

Finally, the texts in each author-based corpus should also range across as many subjects as possible. This additional requirement is necessary because all of the attribution algorithms being tested here are sensitive, to some degree, to a text's meaning. A pure meaning-based attribution algorithm could conceivably be very successful, but only if it picks up on the types of subtle semantic patterns that are a result of an author's worldview and which could therefore be consistent across an author's texts, regardless of subject. For example, the frequency of adjectives may be a useful indicator of authorship because its value depends on how likely an author is to provide extra information about a noun. This is not a stylistic feature of an utterance—the difference between the sentences *the dog barks* and *the black dog barks* is meaningful—but it may still be a useful quantitative indicator of authorship. On the other hand, the frequency of the word *dog* is unlikely to be a good indicator of authorship because it is unlikely that there will be any true consistency in an author's use of this word—except in a poorly designed author-based corpus. If the results of the tests being conducted here are to reflect how well the algorithms can be expected to attribute texts about any subject, then the texts in each author-based corpus must discuss various subjects, otherwise an algorithm capable of topic-based text classification may appear to be capable of author-based text classification. Fortunately, columnists usually write about a wide and ever-changing range of topics, depending on whatever happens to be in the news that day.

4.2.3 Corpus of Possible Authors Compilation

To test the general performance of an attribution algorithm it is not enough for the investigator to assemble a set of highly representative author-based corpora: the collection of author-based corpora must also constitute a highly representative corpus in and of itself. The more representative that this *corpus of possible authors* is, the more realistic and challenging the test that it will provide for an attribution algorithm.

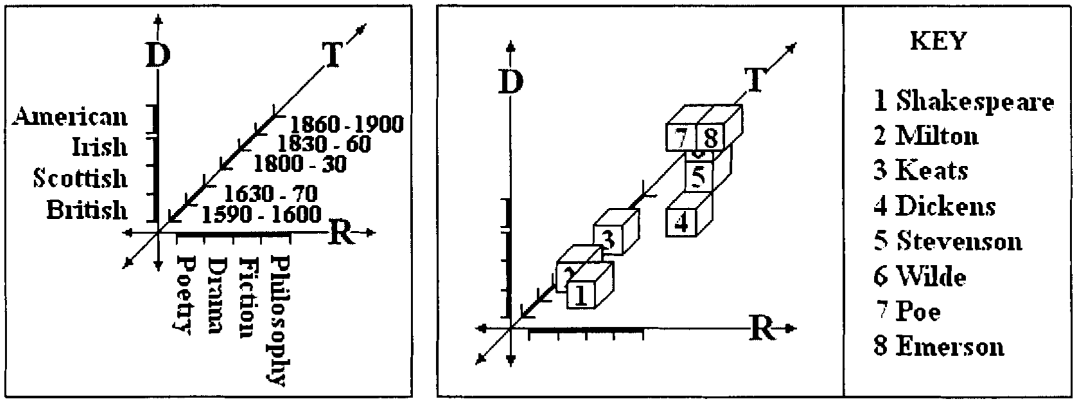
As discussed above, in an actual case of disputed authorship, the dialect, era and register of the anonymous text should direct the compilation of the author-based corpora; therefore the set of author-based corpora will naturally constitute a highly representative corpus of possible authors. For example, if one is attempting to attribute an anonymous eighteenth century poem written in Scots English, then the possible authors should all be eighteenth century Scottish poets; and thus the set of author-based corpora should be a good representation of eighteenth century Scottish poetry. But, in this study, where there is no anonymous text to direct the compilation of the corpus of possible authors, the set of author-based corpora must still constitute a highly representative corpus of possible authors if accurate results are to be obtained.

There are many problems that arise when one attempts to evaluate the performance of an attribution algorithm using an unrepresentative corpus of possible authors, no matter how representative the individual author-based corpora are. For example, consider the set of possible authors used in Peng *et al.* (2003):

William Shakespeare (English Dramatist & Poet, 1590-1620)
John Milton (English Poet, 1630-1670)
John Keats (English Poet, 1815-1825)
Charles Dickens (English Novelist, 1830-1860)
Robert L. Stevenson (Scottish Novelist & Poet, 1870-1900)
Oscar Wilde (Irish Novelist, Dramatist & Poet, 1870-1900)
Edgar Allan Poe (American Short Story Writer, Poet & Critic, 1830-1860)
Ralph W. Emerson (American Philosopher & Poet, 1820-1880)

This diverse set of eight possible authors spans at least four major dialects of English (British, Scottish, Irish, and American), four major written registers (Poetry, Fiction, Drama, and Philosophy), and over three hundred years. The composition of this corpus of possible authors is depicted in Figure 5.

FIGURE 5 CORPUS OF POSSIBLE AUTHORS, PENG *ET AL.* (2003)

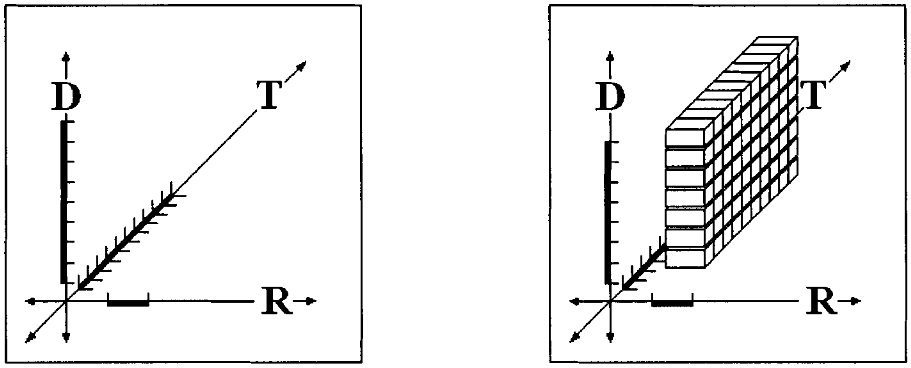


The basic problem with this set of possible authors is that it is unrealistic: it is hard to envision an anonymous text that would inspire an investigator to consider such a small yet such diverse set of possible authors. For example, if the investigator suspects that Shakespeare might have written the anonymous text, then he would certainly analyze the works of Marlowe and Bacon before Edgar Allan Poe. Verisimilitude demands that all the possible authors write in similar varieties of language. But, more important still, a diverse set of possible authors prevents the investigator from accurately gauging the attribution algorithm's performance. For example, it is much more difficult to differentiate between eight Elizabethan poets, than between eight writers of different genres who are spread out over three centuries and the English speaking world. If the test is to be sufficiently challenging, then the possible authors must all write in very similar varieties of language. Furthermore, if an attribution algorithm succeeds over such a stylistically diverse set of possible authors, then the investigator cannot be sure that the algorithm is identifying the personal varieties of language in which each of the possible authors write, or if it is identifying the much larger varieties of language in which each of the possible authors write, as well as many other authors, who happen not to be included in the corpus. If these other authors were included in the corpus, then the algorithm would probably not fare as well. For example, in Peng *et al.*, an algorithm capable of identifying the language of Irishmen would appear to be capable of attributing the texts of

Oscar Wilde, as he is the only Irishman in the corpus; but if the corpus of possible authors included other Irish writers, then this algorithm would fail. The investigator can only know if the attribution algorithm is truly successful if it is tested on a highly representative corpus of possible authors.

To avoid these problems, I have endeavored to compile an ideal corpus of possible authors, which, as illustrated in Figure 6, consists of a set of author-based corpora defined in terms of a single register, the same short span of time, and the most similar dialects.

Figure 6 The Ideal Corpus of Possible Authors



This was accomplished by compiling author-based corpora that represent very similar varieties of language, so that when they are combined they will be highly representative of a slightly larger variety of language. Specifically, the corpus of possible authors used in this study represents the variety of language that is defined in terms of an essentially indivisible register (*Telegraph* opinion column), an era spanning just five years (2000 to 2005), and a very narrowly defined dialect spanning forty possible authors, who all write for basically the same audience (the readership of the London *Telegraph*'s opinion section), and who are mostly from similar social backgrounds (middle-aged, conservative, Anglo-Saxon, middle to upper-class, well-educated, British, Londoner). Overall, I have been least successful in controlling the dialect dimension, in particular the social backgrounds of the possible authors, for while most of the columnists are middle-aged well-educated Britons, some of the columnists—such as Barbara Amiel who is a Canadian, Zoe Heller who lives in New York, and W. F. Deeds who is in his nineties—

are from different social backgrounds. Unfortunately, stricter control of the dialect dimension of the corpus of possible authors would have made it impossible to gather a sufficient number of authors. Nonetheless, this is the most representative corpus of possible authors that has ever been used to test an attribution algorithm.

It is also important that the corpus of possible authors contains a large number of possible authors. First, a large corpus of possible authors allows for the attribution algorithms to be tested on a large number of possible authors simultaneously. For example, most methods have never been asked to attribute a text to one of twenty or to one of forty possible authors. Second, a large corpus of possible authors allows for an attribution algorithm to be tested on multiple smaller sets of possible authors, thereby increasing the accuracy of the test. For example, if one wants to know how well an attribution algorithm can be expected to perform, in general, when asked to attribute a text to one of ten possible authors, then the algorithm should be tested on multiple sets of ten possible authors. For these two reasons, it is important that the corpus of possible authors includes many authors. In this study, the corpus of possible authors contains forty possible authors. A larger set of possible authors was not compiled because of time—both in terms of the compilation of the corpus and of the testing of the algorithms—and because a larger corpus of possible authors would not have allowed for such a highly controlled corpus of possible authors.[96]

Finally, the corpus of possible authors should also be controlled for subject: if the corpus of possible authors is to provide a valid test, then each possible author must write about a similar range of topics. While this is the opposite of the meaning restriction that was placed on the author-based corpora, it serves the same basic purpose: because the attribution algorithms being tested here are sensitive to a text's meaning, it is necessary to ensure that these algorithms are succeeding because they are sensitive to the types of thematic patterns that might be consistent across an author's texts, and not to the types of thematic patterns that are only consistent across an author's texts that are about the same subject. If each possible author wrote about a unique topic, then it would be impossible to know if a successful attribution algorithm was identifying authors or topics. This is not a

[96] Furthermore, none of the methods will prove to be particularly accurate when distinguishing between more than twenty possible authors, and so there is even less reason to compile an even larger corpus.

problem in this study because newspaper opinion columnists, especially when writing at the same time and in the same city, will tend to write about a similar range of subjects.

In Table 3, I present the corpus of possible authors used in this study.

TABLE 3 THE TELEGRAPH COLUMNIST CORPUS

	CODE	NAME	DATE	WORD	SUBJECT
1	AMIE	Barbara Amiel	May 03 – May 04	47,715	WPCR
2	BROW	Craig Brown	Jul 04 – Jan 05	37,860	CPAW
3	CHAN	Alexander Chancellor	Apr 03 – Nov 04	40,844	CPAWR
4	CLAR	Ross Clark	Feb 04 – Jan 05	31,197	EPWCS
5	COLL	Neil Collins	May 03 – Nov 04	40,318	PEC
6	DALE	Janet Daley	Jan 04 – Dec 04	39,380	PWCE
7	DALR	Theodore Dalrymple	Apr 01 – Jan 05	38,504	HCPW
8	DANC	Matthew d'Ancona	Mar 04 – Jan 05	53,891	PW
9	DEED	W.F. Deeds	Jan 04 – Dec 04	26,300	PWCS
10	FARN	Nigel Farndale	Jan 04 – Jan 05	33,024	PCWAS
11	HELL	Zoe Heller	May 03 – Jul 04	41,893	CPW
12	HERB	Susannah Herbert	Oct 02 – Jan 05	36,900	PWC
13	HOWS	Christopher Howse	Jun 04 – Jan 05	30,218	RCAPW
14	IANN	Armando Iannucci	Dec 02 – Jul 04	31,445	PWC
15	JOHB	Boris Johnson	Mar 04 – Jan 05	41,221	PCWE
16	JOHD	Daniel Johnson	Oct 01 – Nov 04	41,078	CPWARE
17	JOHF	Frank Johnson	Nov 03 – Jan 05	24,750	PCW
18	KEEG	John Keegan	May 02 – Jan 05	44,028	WPC
19	LEIT	Sam Leith	May 04 – Dec 04	26,605	CPASW
20	LEWI	Jemima Lewis	Mar 04 – Jan 05	36,311	CWPA
21	MARR	Andrew Marr	Dec 03 – Dec 04	26,798	PWEC
22	MCCA	Jenny McCartney	Mar 04 – Jan 05	37,472	CPAW
23	MOOR	Charles Moore	Sep 03 – Jan 05	48,684	PWC
24	MOUN	Harry Mount	May 02 – Jan 05	31,472	CAPW
25	MYER	Kevin Myers	Mar 04 – Jan 05	38,463	CAPWS
26	NICO	Adam Nicolson	Dec 03 – Jan 05	38,510	CWP
27	PALM	Alasdair Palmer	Jun 02 – Jan 05	37,737	WPC
28	POLL	Stephen Pollard	May 01 – Jan 05	35,538	WCP
29	PRIT	Oliver Pritchett	Jul 04 – Dec 04	31,034	CPW
30	ROBA	Anne Robinson	Apr 03 – May 04	40,982	CAPW
31	ROBS	Stephen Robinson	Dec 03 – Dec 04	36,014	CWPS
32	SAND	Sarah Sands	Mar 04 – Jan 05	36,285	CPSWA
33	SIMP	Peter Simple	Oct 03 – Jan 05	33,741	CPS
34	STEY	Mark Steyn	Apr 04 – Dec 04	45,499	CPWE
35	SYLV	Rachel Sylvester	Sep 03 – Jan 05	38,929	PWE
36	THOM	Alice Thompson	Dec 03 – Jan 05	41,598	PWCE
37	TREF	George Trefgarne	Sep 03 – Jan 05	39,328	EPW
38	UTLE	Tom Utley	Mar 04 – Jan 05	43,453	PCWAE
39	WHIT	Jim White	Mar 04 – Dec 04	39,375	CAPS
40	WOOD	Vicki Woods	Feb 04 – Dec 04	27,916	WPC

For each author-based corpus, I present its code, its author's name, the time span over which its texts were written, the total number of words, and the basic subjects that its texts discuss (B: *British Politics*, W: *World Affairs*, C: *Culture*, A: *Art*, S: *Sport*, E: *Economics*, R: *Religion*, H: *Health*). In total the *Telegraph Columnist Corpus* contains forty authors, 1600 individual texts, and 1.5 million words.

4.3 ATTRIBUTION ALGORITHM EVALUATION

Once the corpus of possible authors has been compiled, it may be used to test the performance of the attribution algorithms. The basic testing procedure is to remove one test text from the corpus of possible authors (i.e. the anonymous text), and to have the attribution algorithm attribute this text by comparing it to the remaining texts in each possible author's corpus (i.e. the writing samples). The test text is then returned to its author-based corpus and the procedure is repeated with a new test text. Once all the texts in the corpus of possible authors have been attributed (forty texts per possible authors), the attribution algorithm's success rate is then calculated by dividing the number of successful attributions by the total number of attributions attempted (forty times the number of possible authors).

In this study, each attribution algorithm is subjected to seven tests. These tests all conform to the basic testing procedure described above, but the number of possible authors included in the corpus of possible authors varies from one test to the next. In particular, the accuracy of each algorithm is tested on seven different sized corpora of possible authors, which contain forty, twenty, ten, five, four, three, and two possible authors. Except for the test involving the full set of forty possible authors, not all the possible authors will be used in any one running of the tests, and therefore it is possible to test the attribution algorithms on multiple sets of possible authors, so that more accurate results may be obtained.[97] Specifically, all the tests conducted in this study that involve fewer than forty possible authors were repeated two-hundred times, using two-hundred different sets of possible authors drawn randomly from the complete set of forty possible authors. This particular number of permutations was arrived at by subjecting various algorithms to tests differing only in the value of this one parameter: it was found that two hundred permutation tests yielded results within 0.5% of the results of one thousand and two thousand permutations tests. When an algorithm is tested over multiple permutations

[97] For example, there are 780 possible ways of choosing 2 authors from a set of 40 possible authors.

of possible authors, its overall success rate is calculated by averaging its success rates over each permutation. In order to ensure that the results of these multiple-permutation tests are commensurable, the same randomly generated sets of possible authors are used every time an algorithm is subjected to these tests.

4.4 SUMMARY

While the corpus of possible authors described in this chapter is not perfect, it is one of the largest and certainly the most representative corpus of possible authors that has ever been used to test the accuracy of an attribution algorithm. This corpus of possible authors should therefore allow for a legitimate test of the general performance of the attribution algorithms introduced in Chapter 3. The results of these tests are presented in Chapter 5.

5 RESULTS & DISCUSSION

5.1 INTRODUCTION

In this chapter, I present the results of testing the attribution algorithms described in Chapter 3, in the manner and on the corpus of possible authors described in Chapter 4. In each of this chapter's sections, I present the results of a subset of these tests, and discuss their significance to the field of quantitative authorship attribution. Finally, I demonstrate that the best results of all are achieved when the values of many textual measurements are considered at the same time. Based on these results I propose a general quantitative approach for resolving cases of disputed authorship.

5.2 ON THE PRESENTATION AND INTERPRETATION OF THE RESULTS

All of the results presented in this chapter take the form of attribution algorithm accuracy tables. Each table presents the results of subjecting multiple attribution algorithms, which are each defined in terms of a set of textual measurements, to multiple tests, which are each defined in terms of the number of possible authors per permutation. The result of subjecting a particular attribution algorithm to a particular test is recorded in the cell at the intersection of the test's row and the algorithm's column, as the percentage of texts correctly attributed.[98] Each algorithm is also numbered for easy look up in the Appendix.

[98] Where percentage of texts correctly attributed is calculated by dividing the total number of texts attributed correctly, by the total number of attempted attributions. For any given test, the number of attempted attributions can be calculated by multiplying the number of texts per author (40), by the number of authors per permutation (2, 3, 4, 5, 10, 20, 40), by the number of permutations (1 for 40 possible authors, and 200 for 2, 3, 4, 5, 10 and 20 possible authors). For example, the results of the 10 possible author test are based on (40 x 10 x 200 =) 80,000 total attempted attributions. Of course, all the texts are attributed multiple times, but each time the text is attributed to a different set of possible authors.

Before these results are presented and discussed, it should first be noted that all the attribution algorithms, in all their forms, have achieved better than random success in all the tests conducted in this study. That is to say, given a test involving *n*-number of possible authors per permutation, where each possible author contributes the same number of texts to be attributed (as in all the tests conducted in this study), a random attribution algorithm, such as an *n*-sided die with each possible author's name written on one side, would correctly assign these texts to their actual authors one out of *n*-times. All the attribution algorithms tested here better this absolute baseline.

Otherwise, there is significant variation in the performance of the attribution algorithms, and it is not altogether clear how to judge their performance. If we are only interested in drawing conclusions about the relative performance of the attribution algorithms, then there is no problem: the algorithms that attribute the highest percentage of texts correctly are the most accurate. But if we are also interested in determining whether any of the attribution algorithms would be useful for resolving an actual case of disputed authorship, then we must also decide how accurate is accurate enough. Such a judgment requires that an arbitrary baseline be set. In this study, if a particular attribution algorithm achieves at least 75% accuracy on a particular test, then that algorithm will be deemed to have performed successfully on that test. This baseline was chosen because, based on my experience and opinion, if an attribution algorithm achieves at least 75% accuracy on a particular test, then that algorithm can reasonably be expected to be of use to investigators attempting to resolve *most* cases of disputed authorship involving as many possible authors or fewer. The reader is free to consider the data in light of a higher or lower baseline if he so chooses.

However, no matter how well an attribution algorithm performs in this study, when an investigator attempts to resolve a case of disputed authorship, he must always retest the algorithm to make sure that it can distinguish between the known works of that particular set of possible authors. If the algorithm proves incapable of distinguishing between that set of possible authors, then it should not be used to help resolve that particular case of disputed authorship. This requirement does not devalue the results of this study: the purpose here is to identify the most *generally* applicable attribution algorithms. The algorithms that prove to be the most successful in this study will not be the best algorithms for distinguishing between *any* set of possible authors, but they should be the best algorithms for distinguishing between *most* sets of possible authors.

5.3 Word- & Sentence-Length

In Table 4, I present the results of testing the attribution algorithms that are based on measurements of word- and sentence-length.

TABLE 4 WORD- AND SENTENCE-LENGTH RESULTS

	TEXTUAL MEASUREMENT			TEST ACCURACY (%)						
	TYPE	VARIANT		POSSIBLE AUTHORS						
		UNIT	RANGE	40	20	10	5	4	3	2
1	Average word-length	Grapheme		7	12	22	39	46	55	70
2	Average sentence-length	Word		6	11	21	37	44	53	69
3	Average sentence-length	Grapheme		6	12	22	39	45	53	70
4	Word-length profile	1 grapheme	1-15 characters	18	26	39	54	60	68	79
5	Word-length profile	1 grapheme	1-10 characters	16	25	37	53	60	68	79
6	Word-length profile	1 grapheme	1-5 characters	11	18	29	45	51	60	74
7	Sentence-length profile	5 words	1-50 words	11	18	29	44	51	60	74
8	Sentence-length profile	5 words	1-30 words	8	16	26	41	47	57	71
9	Sentence-length profile	10 words	1-50 words	10	17	28	44	50	59	73
10	Sentence-length profile	10 words	1-30 words	8	14	24	38	45	54	70
11	Sentence-length profile	25 characters	1-300 characters	12	20	31	46	53	62	74
12	Sentence-length profile	25 characters	1-200 characters	10	17	28	43	50	59	73
13	Sentence-length profile	50 characters	1-300 characters	11	19	30	45	52	61	74
14	Sentence-length profile	50 characters	1-200 characters	9	16	26	41	48	57	72

The first three algorithms tested here, which are based on the value of a single textual measurement of average word- or sentence-length, would appear to be of very little use to investigators of authorship. None of these univariate attribution algorithms have even achieved acceptable results when asked to distinguish between two possible authors. However, the multivariate algorithms did not perform much better: only the larger variants of the multivariate word-length algorithm proved to be capable of distinguishing between two possible authors with any degree of success. The multivariate sentence-length algorithms were not as successful, but of the two types tested here, those that measure sentence-length in characters were slightly more successful than those that measure sentence-length in words. This is not a surprising result considering that only the character-based measurements are sensitive to word-length as well as sentence-length. As was the case with the variants of the multivariate word-length algorithms, the variants of the multivariate sentence-length algorithm based on larger profiles were more successful than the variants based on smaller profiles.

There may be many reasons why the word-length algorithms have achieved better results than the sentence-length algorithms, but perhaps the most important explanation is that a text is always composed of far more words than sentences, and hence any measurement of word-length will be based on far more observations than any measurement of sentence-length. When attributing newspaper articles, which tend to be relatively short, this problem is probably amplified. More thorough explanations are of little practical significance, because both of these approaches to quantitative authorship attribution would seem to be of very limited use in attribution studies. Overall, measurements of word- and sentence-length are poor indicators of authorship.

5.4 Vocabulary Richness

In Table 5, I present the results of testing the attribution algorithms that are based on measurements of vocabulary richness.

Table 5 Vocabulary Richness Results

		Test Accuracy (%)						
		Possible Authors						
	Textual Measurement	**40**	**20**	**10**	**5**	**4**	**3**	**2**
15	Unrestricted Type Token Ratio	8	16	27	44	51	61	75
16	Restricted Type Token Ratio	3	7	14	27	33	42	59
17	Yule's K and Simpson's D	6	10	18	33	38	49	65
18	Guiraud's R	7	13	24	41	48	58	73
19	Herdan's C	7	14	25	42	49	59	73
20	Dugast's k	8	14	24	41	48	56	72
21	Honoré's H	7	13	23	38	45	54	70
22	Sichel's S and Michéa's M	4	9	16	29	35	45	61
23	Entropy	8	14	24	40	47	56	72
24	Tuldava's LN	11	18	31	49	55	64	77
25	W (a = - 0.165)	11	17	26	40	46	53	68
26	W (a = - 0.168)	11	17	26	40	45	52	68
27	W (a = - 0.172)	11	17	26	40	45	52	67

The first two algorithms tested here are the two version of the Type-Token Ratio, where the unrestricted version is calculated for the entire text, and the restricted version is calculated for an equal number of words from each text. Because the shortest text in the entire corpus is 119 words long, the restricted Type-Token Ratio was calculated for only the first 119 words of each text. The unrestricted method clearly outperforms the restricted method, and is, in fact, the second most successful of all the vocabulary richness measures tested in this study, but still it only achieves acceptable results when asked to distinguish between two possible authors.

Yule's *K* and Simpson's *D* (which are functionally equivalent) are far less successful. The relatively poor performance of these two measurements of vocabulary richness is noteworthy because they are the only measurements that Tweedie and Baayen (1998) found to be theoretically stable across texts of different lengths. However, it is not altogether surprising that the theoretically unstable unrestricted Type-Token Ratio has outperformed the theoretically stable *K* and *D,* and the restricted Type-Token Ratio: newspaper columnists tend to write articles of a relatively stable length and therefore text-length itself is a mediocre indicator of authorship over this corpus of possible authors. When text-length was tested *post hoc* as an indicator of authorship, the algorithm achieved 77%, 65%, 56%, 50%, 33%, 20% and 11%, when asked to distinguish between 2, 3, 4, 5, 10, 20 and 40 possible authors, respectively. These results are better than any of the individual vocabulary richness algorithms.

Of the remaining vocabulary richness measurements, Sichel's S and Michéa's *M* (which are reciprocal and thus functionally equivalent), and Honoré's *H,* all perform relatively poorly. While, Entropy and *W* and the various logarithmic attempts to stabilize the Type-Token Ratio—i.e. Herdan's *C,* Guiraud's *R*, Dugast's *k,* Tuldava's *LN*—all perform relatively well. Indeed, the most successful of all the vocabulary richness measurements is Tuldava's *LN*, which is a particularly complex logarithmic manipulation of the Type-Token Ratio. Based on these results, it would appear that *LN* could be of limited use to investigators of authorship.

Overall, measurements of vocabulary richness based on the entire (i.e. *K, D*) or part (i.e. *S, M, H*) of the grouped word frequency distribution have been less accurate than measurements based solely on the number of word-tokens (*N*) and the number of word-types (*V*) in a text (i.e. *TTR, LN, C, R, k, W*). However, none of these algorithms are very successful, probably because they are all based on the values of a single measurement, and because they are all far too sensitive to a text's basic subject matter for this single value to remain constant across an author's texts when those texts range across many different topics. It would thus appear that measures of vocabulary richness are, in general, poor indicators of authorship.

5.5 GRAPHEMES

In Table 6, I present the results of testing the attribution algorithms that are based on measurements of the relative frequency graphemes.

TABLE 6 GRAPHEME RELATIVE FREQUENCY RESULTS

			TEST ACCURACY (%)						
	TEXTUAL MEASUREMENT		POSSIBLE AUTHORS						
	TYPE	VARIANT	40	20	10	5	4	3	2
28	Grapheme profile		25	35	47	62	67	74	83
29	Single-position grapheme profile	1st grapheme in word	20	30	41	56	62	69	80
30	Single-position grapheme profile	2nd grapheme in word	20	29	41	56	62	69	80
31	Single-position grapheme profile	3rd grapheme in word	16	24	35	49	55	63	75
32	Single-position grapheme profile	Last grapheme in word	27	36	49	63	68	73	84
33	Single-position grapheme profile	2nd to last graph in word	23	31	43	57	63	70	81
34	Single-position grapheme profile	3rd to last graph in word	19	28	41	56	61	69	80
35	Multi -position grapheme profile	1st 3 graphemes in word	34	44	56	69	73	79	87
36	Multi -position grapheme profile	1st 6 graphemes in word	43	53	64	76	79	84	90
37	Multi -position grapheme profile	Last 3 graphs in word	31	41	53	67	72	77	86
38	Multi -position grapheme profile	Last 6 graphs in word	42	52	63	74	79	83	90
39	Multi -position grapheme profile	First & last 6 graphs	49	58	68	79	82	86	92
40	Word-internal grapheme profile		28	39	51	65	70	76	85

The first algorithm tested here, based on the simple grapheme profile, is more successful than any of the sentence-length, word-length and vocabulary richness algorithms tested thus far. Nonetheless, like the most successful variants of all these algorithms, the basic grapheme algorithm cannot distinguish successfully between more than two possible authors. The various forms of the single-position grapheme algorithm are even less accurate than the simple grapheme algorithm, except for the word-final grapheme algorithm. Presumably, the relative success of this method is due in part to its sensitivity to an author's use of suffixes. This suggests that the frequency of suffixes could be a good indicator of authorship, though I am aware of no attribution study that has directly examined the frequency of suffixes.

The next set of results presented in Table 6 demonstrate that the performance of the single-position grapheme algorithms can be improved if they are combined to produce multi-position grapheme algorithms. Once again, it would seem that the more measurements included in the profiles, the better the results: the algorithms based on the frequency of graphemes in six positions perform better than the algorithms based on the frequency of graphemes in three positions; and when these two six-position profiles are combined to form one twelve-position profile, the results are even better. For the first time in this study, the six- and twelve-position algorithms achieve acceptable results when asked to distinguish between up to five possible authors. None of these multi-position algorithms have ever been tested before, but it would appear that they are a powerful new tool for investigators of authorship. The final grapheme-based algorithm tested here is the word-internal grapheme profile, which also performs relatively well, successfully distinguishing between up to three possible authors.

The relative success of the grapheme algorithms is likely due to a combination of factors, including those that have been cited in the past, such as an author's preference for particular spellings, synonyms, affixes and words of particular etymological origins. Additional explanations also come to mind, such as an author's aesthetic preferences and—to update Yule's etymological explanation—an author's geographical background. For example, an English writer from Arizona is more likely to use Spanish borrowings than an English writer from Nova Scotia, who would be more likely to use French borrowings. This might cause a difference in the relative frequency of these two authors' use of the grapheme *O* which is more common in Spanish than in French. But perhaps the most important property of graphemes is that they are the most frequent potential indicator of authorship in any English text, and as such any patterns in their usage will have a good chance to emerge. Overall, grapheme frequencies appear to be decent indicators if authorship.

5.6 Words

In Table 7, I present the results of testing the attribution algorithms that are based on measurements of the relative frequency of words. The first eight attribution algorithms tested here are all based on variants of the basic word frequency profile. The largest word profile is the 2-limit profile, which consists of the 265 words that occur in at least 2 of each possible author's texts. All the other word frequency profiles are subsets of these 265 words, where with each successive raising of the limit, a smaller and more frequent set of words remains. The smallest word profile is the forty-limit word profile, which contains only those five words (*and, the, to, a, of*) that occur in every text in the corpus of possible authors. As the limit is raised and the profile shrinks, the content words are the first to be lost. For example, the only content words left in the ten-limit profile are *made, said, time* and *people*, whereas roughly half of the original 265 words are content words.

Table 7 Word Relative Frequency Results

			Test Accuracy (%)						
	Textual Measurement		Possible Authors						
	Type	**Limit**	**40**	**20**	**10**	**5**	**4**	**3**	**2**
41	Word profile	In at least 2 texts per author	44	53	63	73	77	82	88
42	Word profile	In at least 5 texts per author	48	57	67	77	80	85	88
43	Word profile	In at least 10 texts per author	45	54	64	75	79	84	90
44	Word profile	In at least 15 texts per author	40	50	61	73	77	81	88
45	Word profile	In at least 20 texts per author	39	48	59	71	75	80	88
46	Word profile	In at least 25 texts per author	36	46	58	70	74	80	87
47	Word profile	In at least 30 texts per author	33	44	56	70	74	79	87
48	Word profile	In at least 40 texts per author	16	23	35	50	57	64	57

The most accurate word frequency algorithms are based on the five- and ten-limit word profiles. These two profiles still contain most of the function words, but most of the content words have been stripped away. The performance of these two variants is similar, except that the ten-limit variant performs slightly better on sets of two authors, whereas the five-limit variant performs slightly better on all larger sets of possible authors. Both variants successfully distinguish between up to five possible authors. After this point the performance of the algorithm begins to fall off, as function words, as opposed to content words, are removed from the profiles. One of the most common assumptions in quantitative authorship attribution therefore appears to be true: function words are better indicators of authorship than content words. For this reason, unlike most of the other multivariate attribution algorithms tested in this study, larger word profiles do not lead to better results: the largest two-limit variant does not perform as well as the smaller five-limit and ten-limit variants, presumably because of a higher percentage of content words.

It is also important to acknowledge that ignoring the Chi-square frequency restriction has resulted in a better performing algorithm: the two-, five- and ten-limit variants, which contain many words whose frequencies would have been too low to have been considered if the Chi-square test was used conservatively, have outperformed the thirty- and forty-limit variants, whose word frequencies are in the traditional range of the Chi-square test. Therefore not only can the Chi-square restrictions be ignored for theoretical reasons (see section 3.5), but these restrictions should also be ignored for empirical reasons: conforming to these restrictions has led to less accurate results.

Overall, function word frequency has proven to be a good indicator of authorship: these algorithms are some of the most successful techniques tested in this study.

5.7 Punctuation

In Table 8, I present the results of testing the attribution algorithms that are based on measurements of the relative frequency of punctuation marks.

TABLE 8 PUNCTUATION MARK RELATIVE FREQUENCY RESULTS

	TEXTUAL MEASUREMENT		TEST ACCURACY (%)						
			POSSIBLE AUTHORS						
	TYPE	VARIANT	40	20	10	5	4	3	2
49	Punctuation mark profile	By punctuation marks	30	40	53	67	71	77	86
50	Punctuation mark profile	By words	34	45	57	71	75	80	88
51	Punctuation mark profile	By characters	34	46	58	72	76	80	89
52	Grapheme & punct. profile		50	60	70	81	84	87	93
53	Word & punctuation profile	In at least 5 texts per author	63	72	80	87	89	92	95
54	Word & punctuation profile	In at least 10 texts per author	61	69	77	86	88	91	95
55	Word & punctuation profile	In at least 20 texts per author	57	66	75	75	83	87	94

Three versions of the basic punctuation mark algorithm are tested here, which differ, as outlined in Chapter 3, in how the relative frequency of the eight punctuation marks was calculated. When the relative frequency of each punctuation mark is calculated by dividing its frequency in a text, by the total number of punctuation marks in the text, the method does not fare quite as well as when the relative frequency of each punctuation mark is calculated by dividing its frequency in a text by the total number of words, or the total number of characters in the text. Both of these attribution algorithms successfully distinguish between up to four possible authors.

Punctuation mark frequency therefore appears to be a good indicator of authorship. This may be a surprising result to some investigators, as punctuation mark frequency has rarely been analyzed in attribution studies. For this reason, the success of the punctuation algorithms is one the most significant results of this study. These results are particularly impressive because there are only 8 punctuation marks included in the punctuation profile, as opposed to, for example, the 264 graphemes-position pairs included in the largest multi-position grapheme profile. Overall, the frequency of individual punctuation marks is therefore one of the most potent quantitative indicators of authorship.

The similar performance of both algorithms suggests that punctuation mark frequencies could be considered together with either grapheme or word frequencies to yield even more characteristic profiles. Indeed, the algorithm based on the punctuation mark and grapheme profile is capable of distinguishing successfully between sets of up to five possible authors and of achieving 92% accuracy when asked to distinguish between sets of two possible authors; and the algorithm based on the punctuation profile and the five-limit variant of the word profile is the single most accurate individual algorithm tested in this entire study, successfully distinguishing between up to ten possible authors and achieving 95% accuracy when asked to distinguish between two possible authors. The inclusion of punctuation marks in the word and grapheme profiles has thus led to significantly better results.

Punctuation mark frequency is probably a good indicator of authorship because there is so much opportunity for variation in usage: an author can reasonably avoid using every punctuation mark save the period and perhaps the comma and question mark. For example, consider the following quotes from two American authors:

Cormac McCarthy (*Blood Meridian* 1985:337): 128 words and 2 punctuation marks (1 period and 1 comma):

> On the plain behind him are the wanderers in search of bones and those who do not search and they move haltingly in the light like mechanism whose movements are monitored with escapement and pallet so that they appear restrained by a prudence or reflectiveness which has no inner reality and they cross in their progress one by one that track of holes that runs to the rim of the visible ground and which seems less the pursuit of some continuance than the verification of a principle, a validation of sequence and causality as if each round and perfect hole owed its existence to the one before it there on that prairie upon which are the bones and the gatherers of bones and those who do not gather.

Hunter S. Thompson (*Fear and Loathing: On the Campaign Trail '72* 1973:41): 108 words and 30 punctuation marks (9 periods, 5 commas, 4 apostrophes, 3 dashes, 2 question marks, 2 open quotation marks, 2 close quotation marks, 1 semicolon, 1 hyphen, 1 ellipses):

> One canned martini. No beer. A purple TV screen. Both elevators jammed in the basement; fifteen empty bathrooms. Seventy-five cents an hour to park in the lot next door. Chaos and madness in the telephone switchboard. Fear in the back rooms, confusion up front, and a spooky vacuum on top—the eighth floor—where Larry O'Brien is supposed to be holding the gig together...what is he doing up there? Nobody knows. They never see him. "Larry travels a lot," one of the speech writers told me. "He's Number One, you know—and when you're Number One you don't have to try so hard, right?"

Because of the freedom and frequency with which punctuation marks can be used, they are well-suited to be indicators authorship, and their frequencies should be analyzed in any quantitative attempt to resolve modern cases of disputed authorship.

5.8 Positional Stylometry

In Table 9, I present the results of testing the attribution algorithms that are based on measurements of the relative frequency of words in particular sentence-positions, and the relative frequency of collocations.

Table 9 Positional Stylometry Results

	Textual Measurement		Test Accuracy (%)						
			Possible Authors						
	Type	Variant	40	20	10	5	4	3	2
56	Single-position word profile	1st word in sentence	17	30	36	50	56	64	75
57	Single-position word profile	2nd word in sentence	11	18	27	41	47	56	69
58	Single-position word profile	3rd word in sentence	7	13	21	35	41	50	64
59	Single-position word profile	4th word in sentence	6	10	17	30	35	45	59
60	Single-position word profile	Last word in sentence	4	7	13	25	30	39	56
61	Single-position word profile	2nd to last word in sentence	6	11	18	31	37	46	61
62	Single-position word profile	3rd to last word in sentence	6	10	17	29	35	43	59
63	Single-position word profile	4th to last word in sentence	7	11	19	31	36	45	60
64	Multi-position word profile	First 4 words in sentence	22	31	41	55	60	67	77
65	Multi-position word profile	First 8 words in sentence	19	27	38	51	57	63	75
66	Multi-position word profile	Last 4 words in sentence	10	15	24	37	43	51	65
67	Multi-position word profile	Last 8 words in sentence	11	16	25	38	43	52	65
68	Collocation profile	2 words	17	24	34	48	54	61	74
69	Collocation profile	3 words	3	6	11	21	27	35	53

The first algorithm tested here is based on the frequency of words that occur at the beginning of a sentence. This is one of the most successful word position algorithms, and yet it only achieves 75% accuracy when distinguishing between 2 possible authors. The other single-position algorithms are even less successful, most especially when words are counted in relationship to the back of the sentence. In some cases these algorithms barely better the absolute random baseline: e.g. the sentence-final algorithm only achieves 56% accuracy when distinguishing between 2 possible authors, whereas a random attribution algorithm would achieve 50% accuracy. To be fair, the failure of this particular algorithm is largely due to the fact that there are only two words (*it, them*) that occur frequently enough at the end of sentences to be included in the profile at all.

When the single-position profiles are combined, the resulting multi-position profiles do not fare much better. The best of these multi-position variants is based on the frequency of words occurring in the first four positions of a text's sentences, but even this algorithm is only slightly more successful than the best single-position algorithm. Overall these sentence-position methods have therefore proven to be quite inaccurate. Like measures of sentence-length, this lack of success is probably because these measurements are based on the frequency of fairly infrequent strings of characters.

The collocation algorithms also did not perform well. In the case of the three-word collocation algorithm—which is the least successful of all the algorithms tested in this study—this lack of success is not surprising, as *one of the* is the only three-word collocation that is frequent enough to be included in the profiles. On the other hand, the failure of the two-word collocation algorithm was unexpected: even though it contains a relatively large number of collocations, and even though individual words have proven to be good indicators of authorship, the 2-word collocation algorithm does not even achieve 75% accuracy when distinguishing between 2 possible authors.

Positional stylometry has often been criticized in the past, and based on these results it would appear that the critics have been justified: most of the algorithms tested here appear to be too inaccurate to be of any use to investigators of authorship.

5.9 N-Grams

In Table 10, I present the results of testing the attribution algorithms that are based on the relative frequency of character-level n-grams.

TABLE 10 N-GRAM RELATIVE FREQUENCY RESULTS

			TEST ACCURACY (%)						
	TEXTUAL MEASUREMENT		POSSIBLE AUTHORS						
	TYPE	LIMIT	40	20	10	5	4	3	2
70	2-gram profile	In at least 2 texts per author	58	69	77	84	86	89	94
71	2-gram profile	In at least 10 texts per author	65	72	79	86	88	91	94
72	2-gram profile	In at least 20 texts per author	60	69	77	85	87	90	94
73	3-gram profile	In at least 2 texts per author	56	68	75	82	85	89	92
74	3-gram profile	In at least 10 texts per author	61	70	78	85	88	91	94
75	3-gram profile	In at least 20 texts per author	61	71	77	85	88	91	94
76	4-gram profile	In at least 2 texts per author	56	64	72	81	84	88	92
77	4-gram profile	In at least 10 texts per author	55	64	73	83	85	89	93
78	4-gram profile	In at least 20 texts per author	49	58	68	78	82	86	91
79	5-gram profile	In at least 2 texts per author	45	54	66	77	80	84	90
80	5-gram profile	In at least 10 texts per author	47	55	66	76	79	84	90
81	5-gram profile	In at least 20 texts per author	34	43	54	67	71	78	85
82	6-gram profile	In at least 2 texts per author	35	46	57	70	73	78	86
83	6-gram profile	In at least 10 texts per author	35	45	56	68	72	78	86
84	6-gram profile	In at least 20 texts per author	23	31	42	56	61	68	79
85	7-gram profile	In at least 2 texts per author	34	42	45	59	64	69	81
86	7-gram profile	In at least 10 texts per author	19	26	38	52	57	65	75
87	7-gram profile	In at least 20 texts per author	12	19	29	44	49	58	71
88	8-gram profile	In at least 2 texts per author	18	24	36	50	55	62	74
89	8-gram profile	In at least 10 texts per author	9	16	25	40	46	54	68
90	8-gram profile	In at least 20 texts per author	7	12	21	35	41	49	66
91	9-gram profile	In at least 2 texts per author	12	18	28	41	46	55	68
92	9-gram profile	In at least 10 texts per author	6	11	19	32	38	46	62
93	9-gram profile	In at least 20 texts per author	4	8	15	28	33	42	60

Overall, the n-gram algorithms are some of the most successful methods tested in this entire study. The most accurate n-gram algorithms are based on the frequency of sequences of 2 and 3 characters. In particular, the 2- and 3-gram algorithms can distinguish between 2 possible authors with 94% accuracy, and can distinguish successfully between up to 10 possible authors. The two-gram algorithms are slightly more successful, barely outperforming the three-gram algorithms when distinguishing between larger sets of possible authors. From here the performance of the algorithms steadily falls off: the four-, five- and six-gram profiles are fairly good indicators of authorship, whereas the seven-, eight- and nine-gram profiles are weak indicators of authorship. Furthermore, the size of the short n-gram profiles seems to matter very little: all three limits produced very similar results, although the most successful two-, three-, four- and five-gram algorithms are all based on the ten-limit profiles. On the other hand, the size of the longer n-grams profiles is significant: the six-, seven-, eight- and nine-gram algorithms performed best when the size of the profiles was maximized.

The success of the short n-gram algorithms is consistent with the results of the rest of this study: the frequencies of shorter linguistic units have proven to be better indicators of authorship than the frequencies of longer linguistic units, presumably because they are more frequent and because they are less sensitive to the subject of a text. However, this result contradicts past research: Keselj *et al.* (2003), Peng *et al.* (2003) and Clement and Sharp (2003) all achieved their best results using longer n-grams. But there has always been good reason to question the conclusions of these researchers, because long n-grams are known to be good indicators of topic and are often used in topic-based text classification. There is no possible way that any textual measurement can be both a good general indicator of authorship and a good general indicator of subject, because the measurement would be unable to distinguish between text written by different authors on the same subject. Of course, in an attribution study, if each possible author wrote about a unique subject—as was certainly the case in Keselj *et al.* and Peng *et al.*—then a topic-based text classification algorithm would appear to be a good author-based text classification algorithm. Because of the careful experimental design of this study, it is likely that the results obtained here are the more accurate: in general, shorter n-grams are probably better indicators of authorship than longer n-grams.

5.10 Overall Results

In Table 11, I present a ranked list of textual measurements, where only the most successful variants of each basic type are listed.

The first thing to note about the overall results of this study is that some of the quantitative authorship attribution algorithms have proven to be successful, and would still be considered successful even if the arbitrary 75% accuracy limit set at the beginning of this chapter was significantly raised. This is not a trivial result: in the past, critics of quantitative authorship attribution have been justified, to some extent, in questioning the basic assumptions of quantitative authorship attribution because, until now, the methods of quantitative authorship attribution have never been tested together, and never on a corpus of possible authors as large, as realistic and as challenging as the corpus of possible authors used here. These results are proof that the quantitative comparison of texts is a legitimate approach to authorship attribution.

TABLE 11 OVERALL RESULTS

		TEST ACCURACY (%)						
	TEXTUAL MEASUREMENT	**POSSIBLE AUTHORS**						
CODE	**(VARIANT)**	**40**	**20**	**10**	**5**	**4**	**3**	**2**
WPP	Word and punctuation mark profile (5-limit)	63	72	80	87	89	92	95
2NG	2-gram profile (10-limit)	65	72	79	86	88	91	94
3NG	3-gram profile (10-limit)	61	72	78	85	88	91	94
4NG	4-gram profile (10-limit)	55	64	73	83	85	89	93
GPP	Grapheme and punctuation mark profile	50	60	70	81	84	87	93
FBGP	Multi-position graph profile (first & last 6 in word)	49	58	68	79	82	86	92
WP	Word profile (5-limit)	48	57	67	77	80	85	88
5NG	5-gram profile (10-limit)	47	55	66	76	79	84	90
FWGP	Multi-position grapheme profile (first 6 in word)	43	53	64	76	79	84	90
BWGP	Multi-position grapheme profile (last 6 in word)	42	52	63	74	79	83	90
PP	Punctuation mark profile (by character)	34	46	58	72	76	80	89
6NG	6-gram profile (10-limit)	35	45	56	68	72	78	86
WIGP	Word-internal grapheme profile	28	39	51	65	70	76	85
LGP	Single-position grapheme profile (last in word)	27	36	49	63	68	73	84
GP	Grapheme profile	25	35	47	62	67	74	83
7NG	7-gram profile (2-limit)	34	42	45	59	64	69	81
2LGP	Single-position graph profile (2nd to last in word)	23	31	43	57	63	70	81
FGP	Single-position grapheme profile (1st in word)	20	30	41	56	62	69	80
FSWP	Multi-position word profile (first 4 in sentence)	22	31	41	55	60	67	77
WLP	Word-length profile (15 intervals of 1 character)	18	26	39	54	60	68	79
FWP	Single-position word profile (1st word in sentence)	17	30	36	50	56	64	75
8NG	8-gram profile (2-limit)	18	24	36	50	55	62	74
2WC	2-word collocation profile	17	24	34	48	54	61	74
LN	Tuldava's LN	11	18	31	49	55	64	77
SLGP	Sentence-length profile (12 intervals of 25 chars)	12	20	31	46	53	62	74
SLWP	Sentence-length profile. (10 intervals of 5 words)	10	17	28	44	50	59	73
9NG	9-gram profile (2-limit)	12	18	28	41	46	55	68
TTR	Type Token Ratio	8	16	27	44	51	61	75
C	Herdan's C	7	14	25	42	49	59	73
R	Guiraud's R	7	13	24	41	48	58	73
ENT	Entropy	8	14	24	40	47	56	72
SLA	Average word-length	7	12	22	39	46	55	70
SLAG	Average sentence-length (in characters)	6	12	22	39	45	53	70
SLAW	Average sentence-length (in words)	6	11	21	37	44	53	69
KD	Yule's K & Simpson's D	6	10	18	33	38	49	65

But the overall performance of these techniques may also be disappointing to some scholars who have already accepted the validity of quantitative authorship attribution. This is because many of the algorithms tested here do not perform as well as they have in past attribution studies, and none of these algorithms have proven to be capable of distinguishing between sets of twenty and forty possible authors. Therefore it is clear that there are limits to the discriminatory power of our current techniques. These sobering results are probably a product of the highly representative corpus of possible authors that was used in this study, which is far larger and far more challenging than the datasets over which most techniques have been evaluated in the past. These results show that there is still much work to be done by quantitative investigators of authorship.

As for the performance of the individual attribution algorithms, the most successful algorithm tested in this study is based on the word and punctuation mark profile. This method has never been tested before, but on this corpus of possible authors it outperforms all other methods. The only other algorithms that have successfully distinguished between up to ten possible authors are based on two- and three-gram profiles. All of these most successful measurements are similar in that they are sensitive to patterns in an author's use of common words and punctuation marks. It would therefore seems that the frequency of punctuation marks and function words—both which express similar grammatical information—are the best indicators of authorship currently at our disposal. The reason that the word and punctuation mark algorithm has outperformed the n-gram algorithms is probably because the word and punctuation profile is a more direct measurement of these two indicators of authorship: the frequency of an n-gram is more likely to be affected by the frequency of content words and hence by the meaning of a text. For example, the frequency of the 3-gram *and* is mainly determined by the frequency of the function word *and*, but its frequency is also affected by an author's use of such content words as *england* and *landmine* and *andy*—words which are not usually good indicators of authorship. On the other hand, the word and punctuation mark profile is not affected by such thematic patterns. Nonetheless, all three of these methods should probably be applied by quantitative investigators of authorship in all cases of disputed authorship.

Other algorithms that have proven to be capable of distinguishing between up to five possible authors are based on the multi-position grapheme profile, the grapheme and punctuation profile, the word frequency profile, and the four-gram profile. Based on these results, it would appear that these methods also deserve to be applied in most cases of disputed authorship. A number of other algorithms were also found to be of use in smaller cases of disputed authorship. These algorithms are based on the word-internal grapheme profile, the word-initial grapheme profile, the word-final grapheme profile, the basic grapheme profile, the five-, six-, seven- and eight-gram profiles, the word-length profile, the first four words in a sentence profile, the first word in a sentence profile, and two vocabulary richness measures (Tuldava's LN and the Type Token Ratio). It also appears that the two-word collocation profile and the sentence-length profile (in characters) may also be useful indicators of authorship in the occasional case of disputed authorship. Overall, there is thus a fairly large battery of textual measurements that have proven to be of useful indicators of authorship.

These most successful sets of textual measurements share certain properties that should be enumerated here so that they may inform the conception of new and hopefully more characteristic indicators of authorship. First, I reiterate that a textual measurement is defined as a function of the frequencies of one or more strings of characters in a text. It should therefore be no surprise that the most successful sets of textual measurements are based on *frequent* strings of characters: if a string is used by an author at a stable enough rate to be a good indicator of authorship, then it must be used frequently enough for its relative frequency in a text to be a good estimate of its relative frequency in its author's population of texts. A corollary is that shorter strings will tend to be better indicators of authorship than longer strings, because, as Zipf (1949) first noticed, shorter strings are more frequent than longer strings.[99] Indeed, in this study, the algorithms that are based on the frequency of shorter strings (e.g. function words, short n-grams) have generally outperformed the algorithms that are based on the frequency of related longer strings (e.g. content words, collocations, long n-grams).

Second, the most successful sets of textual measurements are based on strings whose usage depends mainly on an author's style. None of the measurements of authorship tested here are wholly stylistic, as all are sensitive to the meaning of a text.[100] For example, measurements of word-length, sentence-length, grapheme frequency, and function word frequency could all distinguish between two texts that are identical except that the name of the main character is *The Secretary of State* in one and *The President* in the other. We can therefore conclude that none of these measurements are purely stylistic, because all are capable of distinguishing between these two stylistically identical texts. This is not necessarily a problem: our measurements could mainly be sensitive to aspects

[99] Shorter strings tend to be more frequent because most strings are composed of sub-strings, many of which reoccur in other strings as well.

[100] Thorndike's measurements of an author's rate of contraction, and perhaps certain word-ratios (e.g. *while / (while + whilst)*) come close to pure stylistic measurements.

of an author's worldview which are consistently expressed across his works. For example, an author's frequent use of the word *at* may reflect his fixation on time, and an author's frequent use of long sentences may reflect his tendency to repeat information. However, while such quantitative thematic evidence could be useful, the results of this study suggest that pure stylistic measurements are more useful, as the most successful methods appear to be those least affected by a text's meaning: function words are better indicators of authorship than content words, individual words are better than collocations, and words and punctuation mark are better than n-grams. In the future, it would therefore appear that investigators of authorship would be wise to focus on developing purely stylistic measurements.

Despite their promise, it would seem that investigators have not focused on purely stylistic measurements in the past either because they were not aware that such measurements exist (most investigators seem to believe that our current measurements are wholly stylistic), or because it is far more difficult to determine their values in a text. For example, the relative frequency of complement clauses marked by *that* is a pure stylistic measurement, because it is based on the frequency of two (semantically equivalent) variants of the same linguistic variable (e.g. *he thinks that the dog is fast* vs. *he thinks the dog is fast*). But in order to determine its value one must count all the complement clauses in a text, and all the complement clauses that are marked by *that*. Even the second half of this procedure is relatively tricky because any *that* acting as a determiner must be ignored (e.g. *he thinks that dog is fast*). There are many other frequently occurring linguistic variables that could also be measured, but all require far more complex procedures than those described here to count such strings as words or n-grams or even sentences. Hopefully as linguistic theories and linguistic parsing and tagging technologies advance, investigators of authorship will begin to exploit such purely stylistic measurements.

Third, the most successful algorithms are based on large profiles that contain many textual measurements. Indeed, the most accurate algorithm was based on a profile containing a combination of punctuation mark and word frequencies. It is important to base an attribution on the values of multiple textual measurements because there is no perfect textual measurement: the value of no textual measurement is absolutely stable across every author's texts and absolutely variable across every set of possible authors. There is simply far too much grammatical and meaningful variation across an author's texts, and far too many authors, for an author's texts to be characterized by the value of a single textual measurement. For this reason, multivariate approaches to quantitative authorship attribution, based on the values of multiple textual measurements, achieve the best results. It would therefore seem that the best results of all could be obtained by taking into consideration the values of as many textual measurements as possible. To test this theory, in the next section, I will conduct a *post hoc* analysis of the results of this study to see if a method based on a combination of the most successful attribution algorithms would have performed better than any of the individual attribution algorithms.

5.11 COMBINATION OF TECHNIQUES

Because the most successful attribution algorithm tested in this study is based on a combination of the word and punctuation mark profiles, it would seem that an even more successful attribution algorithm would be based on an even larger number of textual measurements. But this hypothesis would hold true only if the measurements being combined were having trouble attributing the texts of different possible authors: if all the measurements were failing to attribute the texts of the same problematic subset of possible authors, then the investigator's best choice would be to only use the best measurement, because otherwise the results of the others would only drown its more dependable attributions. However, this does not appear to be the case.

In Table 12, I present the results of an author by author error analysis for a selection of the most successful textual measurements, where the possible authors are ranked from least to most difficult to attribute for each textual measurement. By looking across almost any author's row, it can be verified that the various measurements do in fact have difficulty attributing the texts of different authors: while there are certain authors that easy to identify (e.g. DALE, DANC, JOHB), and other authors that are difficult to identify (e.g. HOWS, IANN, MOUN), overall there are many possible authors that are sometimes easy and sometimes difficult to attribute. It is therefore reasonable to hypothesize that even better results may be achieved by combining the results of many textual measurements.

TABLE 12 ERROR ANALYSIS FOR SELECTED TEXTUAL MEASUREMENTS

	TEXTUAL MEASUREMENTS															
AUTHOR	WPP	2NG	3NG	4NG	GPP	FBG	WP	5NG	PP	WIG	2WC	SLGP	WLP	FSW	LN	TTR
AMIE	32	8	23	14	17	11	23	18	27	9	33	30	14	18	1	23
BROW	36	29	34	1	38	31	19	34	39	32	20	38	34	39	31	33
CHAN	6	6	4	35	28	13	3	10	28	22	3	33	26	3	26	19
CLAR	18	38	32	22	31	37	33	16	17	35	17	6	33	19	32	34
COLL	14	23	11	10	9	26	2	14	10	29	16	14	36	8	11	18
DALE	9	10	6	16	2	10	16	13	9	5	9	9	1	22	10	17
DALR	19	17	2	9	25	16	5	5	11	24	10	8	4	2	33	14
DANC	11	9	3	13	8	14	4	2	19	4	2	7	6	7	3	11
DEED	15	19	1	34	32	23	6	32	22	37	35	37	27	11	9	7
FARN	12	15	26	24	15	17	12	24	35	11	15	20	30	29	15	24
HELL	35	25	27	18	23	27	24	17	20	8	24	24	11	12	20	28
HERB	10	11	25	30	7	18	29	33	2	20	34	11	23	35	23	15
HOWS	38	34	38	39	39	38	40	37	33	39	27	35	40	33	24	29
IANN	37	40	36	29	34	40	37	39	31	38	40	40	38	40	22	38
JOHB	7	21	15	5	4	12	13	6	3	17	4	15	8	6	2	10
JOHD	33	31	31	33	29	34	31	30	36	25	30	36	17	26	34	30
JOHF	1	30	16	31	24	29	20	35	23	27	28	16	29	31	8	21
KEEG	20	13	7	11	5	1	25	4	18	1	21	29	2	20	17	13
LEIT	16	22	37	40	21	25	39	40	12	28	36	19	25	36	4	4
LEWI	17	7	13	25	10	7	22	19	6	13	22	21	19	27	21	3
MARR	2	2	12	21	6	4	21	25	7	15	31	34	16	23	6	1
MCCA	23	16	28	26	18	21	28	36	15	6	38	3	13	38	27	31
MOOR	25	26	21	2	19	32	11	12	16	21	12	23	21	5	5	12
MOUN	4	33	39	36	16	28	27	31	5	30	19	32	32	28	38	40
MYER	34	24	19	27	22	19	30	29	25	14	37	5	10	15	12	22
NICO	29	35	35	19	35	30	26	11	24	34	13	4	28	16	28	20
PALM	13	32	20	7	37	33	18	7	34	26	8	26	20	14	37	9
POLL	26	39	33	38	33	35	34	28	32	33	14	28	31	10	36	32
PRIT	40	37	40	28	40	39	35	26	40	40	18	39	39	30	30	39
ROBA	27	1	5	15	12	2	8	20	29	10	32	25	12	17	18	8
ROBS	39	36	30	20	36	36	38	22	14	31	23	1	15	37	39	37
SAND	30	3	14	23	14	6	36	23	8	12	25	2	7	34	40	35
SIMP	5	5	8	17	3	5	32	21	13	3	29	10	5	25	29	5
STEY	8	12	22	32	13	9	7	27	4	36	26	13	9	1	16	36
SYLV	21	14	10	6	11	15	17	3	21	2	6	18	18	24	35	26
THOM	24	20	18	8	20	20	14	8	26	18	7	22	22	13	13	27
TREF	31	28	24	12	27	22	9	15	30	16	11	17	37	21	14	25
UTLE	22	27	17	4	30	24	1	1	38	23	1	31	24	4	19	2
WHIT	28	18	9	3	26	8	10	9	37	19	5	27	35	9	25	16
WOOD	3	4	29	37	1	3	15	38	1	7	39	12	3	32	7	6

In this section, I test two additional attribution algorithms that take into consideration the results of all sixteen of the individual attribution algorithms listed in Table 12. The first seven methods were chosen because they have achieved at least 75% accuracy when distinguishing between 5 possible authors. These most successful attribution algorithms are based on the word and punctuation profile, the word profile, the grapheme and punctuation profile, the two-, three- and four-gram profiles, and the front and back of word grapheme profile. The remaining nine algorithms were chosen so as to include the results of a wide range of textual measurements. Therefore I have also included algorithms based on the word-length profile in characters, the sentence-length profile in characters, Tuldava's *LN*, the Type-Token Ratio, the word-internal grapheme profile, the punctuation profile, the 5-gram profile, the two-word collocation profile, and the multi-sentence-position word profile.

In order to combine the results of these sixteen attribution algorithms, the sets of textual measurements upon which each is based cannot be combined to make one gigantic textual profile. This is because many of these sets of textual measurements are in different scales, and it would therefore be inappropriate and ineffective to use the Chi-square statistic to compare all their values simultaneously. The approach taken here is therefore to attribute a text by applying each algorithm individually, and by then outputting the author that most of the attribution algorithms have selected.

The results of testing two variants of this combination algorithm are presented here. These variants differ in terms of how many votes are given to each of the individual attribution algorithms: in the simple version, each algorithm is given one vote; in the weighted version, each algorithm is given a number of votes based on its individual success.[101] The results of applying these two version of the combination algorithm to the regular set of seven tests is presented in Table 13, where I have also included the results of the punctuation and word frequency algorithm and the two-gram frequency algorithm for comparison.

TABLE 13 COMBINATION ALGORITHM RESULTS

	TEST ACCURACY (%)						
TEXTUAL MEASUREMENT	POSSIBLE AUTHORS						
(VARIANT)	40	20	10	5	4	3	2
Weighted Combination	69	78	85	91	93	95	97
Simple Combination	58	72	82	90	92	94	96
Word & Punctuation Mark Profile (5-limit)	63	72	80	87	89	92	95
2-gram profile (10-limit)	65	72	79	86	88	91	94

[101] In particular, (algorithm-votes): WPP-4, 3NG-3, 2NG-3, GPP-3, FBGP-2, 4NG-2, WP-2, PP-2, 5NG-2, WIGP-1, WLP-1, LN-1, FSWP-1, 2WC-1, TTR-1, SLGP-1.

The two combination algorithms are the most accurate algorithms tested in this entire study. The simple combination algorithm equaled or bettered the best individual algorithms on six out of the seven tests. It did not perform as well in the forty-author test because most of the sixteen algorithms performed very poorly at this level, and so their votes overwhelmed the votes of the few attribution algorithms that are more successful. But this problem can be overcome by weighing the votes of each of the algorithms: the weighted combination algorithm has performed significantly better than every other algorithm tested in this study on all seven of the tests. Most notably, the weighted combination algorithm is the first algorithm that has successfully distinguished between twenty possible authors, and that has distinguished between five possible authors with over 90% accuracy. Based on these results, it would appear that the best approach to quantitative authorship attribution is one that is based on the results of as many proven attribution algorithms as possible, where the significance of each individual attribution is weighted according to the individual performance of its algorithm.

However, it should be made clear that the evaluation of the two combination algorithms was an unplanned or *post hoc* experiment: the individual algorithms were tested first, and then the most successful algorithms were combined and tested on the same dataset. For this reason it was already very likely that the combination algorithms would outperform the individual algorithms, especially after considering the data presented in Table 12. Nonetheless, this does not weaken the strength of my conclusion, because I have only concluded that when attempting to resolve a case of disputed authorship an investigator should apply a variety of attribution algorithms. I have not concluded that this specific combination of algorithms is the most generally applicable. Indeed, I have not even concluded that this specific combination of algorithms is best for distinguishing between this set of authors, as I have only tested two possible combinations. All that I have demonstrated is that some combination of algorithms will usually outperform any individual algorithm, and therefore that a combination of algorithms should be used to resolve cases of disputed authorship. It is responsibility of the investigator to determine which combination can best distinguish between that particular set of possible authors. This conclusion may seem to be too obvious to be stated, but this argument has rarely been made, and most attribution studies today are only based on a single type of textual measurement.

Based on the results of this study, I therefore propose the following procedure to resolve all cases of disputed authorship. First, the investigator must identify a valid set of possible authors through an analysis of the external evidence of the anonymous text. Second, the investigator must compile a corpus of possible authors by collecting a large sample of each author's writings, which are as stylistically similar as possible to the anonymous text. Third, the investigator should test a wide range of attribution algorithms on the corpus of possible authors so as to establish which algorithms can best distinguish between that particular set of possible authors. Fourth, the investigator should test various weighted combinations of the best algorithms on the same corpus of possible authors. Finally, once an acceptably accurate combination of algorithms has been identified, the investigator may then use these algorithms to compare the anonymous text to each author-based corpus, in order to determine which possible author is the best match.

5.12 SUMMARY

In this chapter, I have presented the results of testing a wide range of attribution algorithms on a large and carefully constructed corpus of possible authors. For the first time in the history of quantitative authorship attribution, investigators now have access to reliable data about which of our textual measurements are the most useful for attributing authorship. In particular, the most successful measurements are based on the frequency of common words, punctuations marks and character-level n-grams. But the best results of all were achieved when the outputs of many attribution algorithms were combined. Based on these results I have proposed a general approach to quantitative authorship attribution, which involves analyzing the values of many different types of textual measurements

6 CONCLUSION

In unselfconscious utterance, certain features occur—relatively permanent features of the speech or writing habits—which identify someone as a specific person, distinguishing him from other users of the same language.

David Crystal & Derek Davy (1969:66)

The anonymous text is open to interpretation. When different people write the same words, we do not necessarily interpret these words in the same way; but if we do not know who wrote the words, then we must allow for many different interpretations. This is why it is important to know the source of a text: by limiting its possible authors we also limit its possible meanings. Knowledge of source simplifies understanding. Knowledge of source allows us to judge the information that a text conveys.

Usually, the physical transmission of a text allows its reader to determine its author: we remember how and where we obtained the document, we recognize handwriting, we check for postmarks and email addresses. When this is not enough we may analyze the information that the text contains, in search of self-references and familiar opinions. But when all else fails a text can also be attributed by comparing the values of a set of textual measurements in the text to their corresponding values in a series of possible author writing samples. This is the basic approach of quantitative authorship attribution.

There are two reasons to develop quantitative authorship attribution algorithms. First, an attribution algorithm can be used to resolve tricky cases of disputed authorship.

For example, if literary or forensic experts do not agree on the provenance of a text, then an attribution algorithm can act as an arbitrator—tested to the satisfaction of all the experts, capable of producing replicate attributions free of human bias and error. The second reason to develop attribution algorithms is that they can be automated. At this time, automation primarily allows for investigators to rigorously test and fine-tune their techniques, but as these techniques improve, automation will also allow for unsupervised attribution applications. For example, a generally applicable attribution algorithm could be used to sort through web-pages and emails based on authorship, or to scan through term papers and confessions in search of plagiarized passages, or to filter through reams of anonymous internet chatter for the words of a flagged author. A generally applicable attribution algorithm could also play an ancillary role in applications that analyze the meaning of texts. For example, a topic-based text classification program would benefit from the ability to search for a particular author's texts when queried about that author or a topic with which that author is associated. And because knowledge of source helps human readers to resolve ambiguity and understand the meaning of texts, an attribution algorithm could also be a useful component in data extraction and machine translation applications.

In order for attribution algorithms to help solve any of these problems, it is necessary to discover sets of textual measurements whose values are both relatively stable across most author's texts, and relatively variable across most sets of possible authors. Over the past two centuries, many different sets of textual measurements have been proposed, but surprisingly, until now, there has never been a large-scale comparison of these various quantitative indicators of authorship. To conduct this experiment, I had to first identify the basic types of textual measurements that have been proposed by investigators of authorship. This proved to be a difficult task because the history of the field is scattered across academia: attribution algorithms have been developed by literary scholars, statisticians, theologians, forensic scientists, linguists, computer scientists, physicists, classicists, philosophers, mathematicians, economists, a political scientist and an entomologist. This diversity is also reflected in the fact that there is no standard testing procedure for attribution algorithms. Therefore I had to develop a rigorous experimental design as well, focusing in particular on how to compile the highly representative corpus of possible authors that is necessary to conduct a proper evaluation of the performance of an attribution algorithm.

Some of the results of this study have confirmed past assumptions. For example, function word frequency—the most commonly used type of textual measurement in modern attribution studies—has proven to be a good indicator of authorship. Other results have been unexpected. For example, collocation frequency and long n-gram frequency—two types textual measurements that have been highly touted in recent research—have proven to be weak indicators of authorship. On the other hand, punctuation mark frequency—an often overlooked type of textual measurement—has proven to be a very useful indicator of authorship, especially when combined with word and grapheme frequencies. Indeed, the most accurate individual algorithm tested in this entire study is based on a novel combination of word and punctuation mark frequency. The short n-gram and the grapheme-position algorithms also performed very well, even though these textual measurements have failed to attract much attention in past research. Overall, it was therefore possible to identify a fairly large collection of useful textual measurements. But based on the results of this study it is also clear that our current measurements are not as accurate as we would like. For example, none of the individual algorithms were even moderately successful when asked to distinguish between sets of twenty and forty possible authors.

These limits were overcome, in this study, at least to some extent, by combining the results of many different attribution algorithms. From a practical standpoint, this is probably the most important result of this study, because at this time most investigators base their attributions on a single type of textual measurement. But the results of this study have also allowed for the properties of useful textual measurements to be identified, so that they may direct the selection of textual measurements in future attribution research. In particular, the most useful sets of textual measurements have tended to be frequent, stylistic, and large. This is a very important result because, as this study has demonstrated, the best approach to quantitative authorship attribution is one that is based on the values of as many textual measurements as possible. Fortunately, there are many untapped measurements that conform to these properties, such as the relative frequency of parts-of-speech, phrases, morphological affixes, discourse markers, syllables, near-collocations, words of particular etymological origins and stress patterns. But, in my opinion, the most promising types of textual measurements are based on the relative frequency of the variants of linguistic variables. These measurements are the true coordinates of style by which we may locate the texts of one author in the universe of language. If the field of quantitative authorship attribution is to reach its full potential then we must begin to make use of these purely stylistic indicators of authorship.

REFERENCES

A.B.M. (1890). Curves of Literary Style. *Science*, 13(320): 226.

Adair, D. (1944a). The authorship of the disputed Federalist papers. Part I. *The William and Mary Quarterly*, 1: 97-122.

Adair, D. (1944b). The authorship of the disputed Federalist papers. Part I. *The William and Mary Quarterly,* 1: 235-264.

Ager, D. E. & Knowles, F. E. & Smith, J. (eds) (1979). *Advances in Computer-Aided Literary and Linguistic Research.* Birmingham: AMLC.

Aitken, A. J. & Bailey R. W. & Hamilton-Smith, N. (eds.) (1973) *The Computer and Literary Studies.* Edinburgh University Press.

Baayen, R. H. (2001). *Word Frequency Distributions.* Dordrecht: Kluwer Academic Publishers.

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: 110-120.

Bailey, Richard W. (1969), Statistics and Style: A Historical Survey. In *Statistics and Style* (Doležel & Bailey eds.): 217-236.

Baker, J.C. (1988). Pace: A Test of Authorship Based on the Rate at Which New Words Enter an Author's Text. *Journal of the Association for Literary and Linguistic Computing,* 3: 36-39.

Bee, R. E. (1971). Statistical Methods in the Study of the Masoretic Text in the Old Testament. *Journal of the Royal Statistical Society A,* 134: 611-622.

Bee, R. E. (1972). A Statistical Study of the Sinai Periscope. *Journal of the Royal Statistical Society A,* 135: 406-421.

Bennett, P. E. (1957). The Statistical Measurement of a Stylistic Trait in *Julius Caesar* and *As You Like It. Shakespeare Quarterly,* 8: 33-50.

Bennett, W. R. (1976). *Scientific and Engineering Problem-Solving with the Computer.* Englewood Cliffs, NJ: Prentice Hall, Inc.

Biber, D. (1994). An Analytical Framework for Register Studies, In Biber, D. & Finegan, E. (eds) *Sociolinguistic Perspective on Register*: 31-55.

Biber, D. (1995). *Dimensions of Register Variation.* Cambridge University Press.

Biber, D. & Finegan, E. (eds) (1994). *Sociolinguistic Perspective on Register.* New York: Oxford University Press.

Biber, D. & Johansson, S. & Leech, G. & Conrad, S. & Finegan, E. (1999). *Longman Grammar of Spoken and Written English.* London: Longman.

Binongo, J. N. G. (1993). Incongruity, Mathematics and Humor in *Joaquinesquerie. Philippine Studies*, 41: 477-511.

Binongo, J. N. G. (1994). Joaquin's *Joaquinesquerie, Joaquinesquerie's* Joaquin: a statistical expression of a Filipino writer's style. *Literary and Linguistic Computing,* 9: 267-279.

Binongo, J. N. G. (1995). *Tropical Gothic* versus *Joaquinesquerie*: quantifying their qualitative differences. *Philippine Studies,* 45: 66-92.

Binongo, J. N. G. & Smith, M.W.A. (1999a). The Application of Principal Component Analysis to Stylometry. *Literary and Linguistic Computing,* 14: 445-465.

Binongo, J. N. G. & Smith, M.W.A. (1999b). A Bridge Between Statistics and Literature: The Graphs of Oscar Wilde's Literary Genres. *Journal of Applied Linguistics,* 26: 781-787.

Bloomfield, L.(1933). *Language.* New York: Holt, Rinehart and Winston.

Boyle, R. (1886). *Transactions of the New Shakespeare Society,* 1880-1886:443-487.

Brinegar, C.S. (1963). Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship. *Journal of the American Statistical Association,* 58:85-96.

Brown, M.P. (1963). *The Authentic Writings of Ignatius: a Study of Linguistic Criteria.* Durham, NC: Duke University Press.

Brunet. E (1978). *Le Vocabulaire de Jean Giraudoux. Structure et Evolution.* Geneva: Slatkine.

Burrows, J. F. (1987).*Computation into Criticism: a Study of Jane Austen's Novels and an Experiment in Method.*. Oxford University Press.

Burrows, J. F. (1992a). Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information. *Literary and Linguistic Computing,* 7: 91-109.

Burrows, J. F. (1992b). Computers and the Study of Literature. In C.S. Butler (ed.), *Computers and Written Texts.* Oxford: Blackwell.

Burrows, J. F. & Craig, H. (1994). Lyrical Drama and the 'Turbid Montebanks': Styles of dialogue in Romantic and Renaissance tragedy. *Computers and the Humanities,* 28: 63-86.

Burrows, J.F. & Craig, H. (2001). Lucy Hutchinson and the Authorship of Two Seventeenth-Century Poems: A computational Approach. *The Seventeenth Century,* 16: 259-282

Burrows, J.F. & Hassall, A.J. (1988). *Anna Boleyn* and the Authenticity of Fielding's Feminine Narrative. *Eighteenth Century Studies,* 21: 427-453.

Butler, C.S. (ed.) (1992). *Computers and Written Texts.* Oxford: Blackwell.

Campbell , G.C. & Corns, T.N. & Hale, J.K. & Holmes, D.I. (1997). The Provenance of 'De Doctrina Christiana'. *Milton Quarterly,* 31: 67-121.

Canter, D. (1992). An Evaluation of the 'Cusum' Stylistic Analysis of Confessions. *Expert Evidence,* 1: 93-99.

Cavner, W. & Trenkle, J. (1994). N-Gram-Based Text Categorization. *Proceedings SDAIR-94.*

Chambers, E. (1930). *William Shakespeare: A Study of Facts and Problems.* Oxford: Clarendon Press.

Chambers, J.K. & Trudgill, P. (1980), *Dialectology.* Cambridge University Press.

Chaski, C. E. (2001). Empirical Evaluation of Language-Based Author Identification Techniques. *Forensic Linguistics,* 8: 1- 65.

Clement, R. & Sharp, D. (2003). Ngram and Bayesian Classification of Documents. *Literary and Linguistic Computing,* 18: 423-447.

Conway, J (1959). *Evidential Documents.* Springfield, IL: Charles C Thomas.

Craig, H. (1992). Authorial Styles and the Frequencies of Very Common Words: Jonson, Shakespeare, and the Additions to 'The Spanish Tragedy'. *Style,* 26:199- 220.

Craig, H. (1999). Authorial Attribution and Computational Linguistics: if you can tell authors apart, have you learned anything about them? *Literary and Linguistic Computing,* 14: 103-113.

Crystal, D. (1987) *The Cambridge Encyclopedia of Language.* Cambridge University Press.

Crystal, D. & Davy, D. (1969). *Investigating English Style.* London: Longman.

Cochran, W. G. (1952). The χ^2 test of goodness of fit. *Annals of Mathematical Statistics,* 23: 315-345.

Cochran, W. G. (1954). Some Methods for Strengthening the Common χ^2 Test. *Biometrics,* 10: 417-451.

D* (1974). *Stremya 'Tikhogo Dona': Zagadki romana* (The Mainstream of *The Quiet Don:* Riddles of the Novel). Paris: YMCA Press.

Damereau, F. J. (1975). The Use of Function Word Frequencies as Indicators of Style. *Computers and the Humanities*, 9: 271-280.

Doležel, L. & Bailey, R. W. (eds) (1969). *Statistics and Style*, New York: American Elsevier Publishing Co.

Dale, R. & Moisl, H. & Somers, H. (eds) (200). *Handbook of Natural Language Processing*. Marcel Dekker.

De Haan, P. & Schils, E. (1993). The Qsum Plot Exposed. *Proceedings of the 14th ICAME Conference.* Amsterdam: Rodopi.

De Morgan, S. (1882) *Memoir of Augustus de Morgan by his Wife Sophia Elizabeth de Morgan With Selections From His Letters.* London: Longmans, Green, and Co.

De Vel, O. & Anderson, A. & Corney, M. & Mohay, G. M. (2001). Mining E-mail Content for Author identification Forensics. *SIGMOD Record,* 30: 55-64.

Dolezel, L. & Bailey, R. W. (eds.) (1969). *Statistics and Style.* New York: Elsevier.

Dugast, D. (1979). *Vocabulaire et Stylistique*. Geneva: Slatkine.

Eddy, H.T. 1887. The Characteristic Curves of Composition. *Science*, March 25, 1887:297.

Eggins, S. & Martin, J.R. (1997). Genres and Registers of Discourse. In van Dijk (ed.) Discourse as Structure and Process:230-256. London: Sage.

Ellegård, A. (1962a). *A Statistical Method for Determining Authorship: 1769-1772.* Gothenburg: Acta Universitatis Gothoburgensis.

Ellegård, A. (1962b). *Who Was Junius?* Stockholm: Almquist and Wiksell.

Elliott, W. & Valenza, R. J. (1996). And Then There Were None: Winnowing The Shakespeare Claimants. *Computers and the Humanities*, 30: 191-245.

Elliott, W. & Valenza, R. J. (1998). The Professor Doth Protest Too Much, Methinks. Computers and the Humanities, 32: 425-490.

Elliott, W. & Valenza, R. J. (2001). Smoking Guns and Silver Bullets: Could John Ford have written the Funeral Elegy? *Literary and Linguistic Computing,* 16: 205-232.

Elliott, W. & Valenza, R. (2002). So Many Hardballs So Few Over the Plate. *Computers and the Humanities*, 36: 455-460.

Ellison, John. (1965). Computers and the Testaments, In *Computers for the Humanities.* New Haven: Yale University Press:72-74.

Erdman, D. & Fogel, E. (Eds.) (1966). *Evidence for Authorship: Essays on Problems of Attribution.* Ithaca, NY: Cornell University Press.

Farringdon, J. M. & Morton, A. Q. (1989). Fielding and the Federalist. *Technical Report. CSC 90/R6.* University of Glasgow.

Farringdon, J. M. & Morton, A.Q. & Baker, M. D. (1996). *Analysing for Authorship.* Cardiff: University of Wales.

Farnham, W. (1916). Colloquial Contractions in Beaumont, Fletcher, Massinger, and Shakespeare as a Test of Authorship. *PMLA,* 31: 326-358.

Ferguson, C.A. (1994). Dialect, Register, and Genre: Working Assumptions About Conventionalization. In Biber, D. & Finegan, E. (eds) *Sociolinguistic Perspective on Register*: 15-30.

Fleay, F.G. (1874a). On Metrical Tests as Applied to Dramatic Poetry. *Transactions of the New Shakespeare Society*, I: 1-39.

Fleay, F.G. (1874b). On the Authorship of "The Taming of the Shrew". *Transactions of the New Shakespeare Society*, 1:85-129.

Fleay, F.G. (1876). *Shakespeare Manual.* London: Macmillan.

Forsyth, R. S. & Holmes, D. (1996). Feature-finding for Text Classification. *Literary and Linguistic Computing,* 11: 163-174.

Forsyth, R. S.& Holmes, D. & Tse, E. (1999). Cicero, Sigonio, and Burrows: Investigating the Authenticity of the *Consolatio. Literary and Linguistic Computing,* 14: 375-400.

Foster, D. (1989). *'Elegy' by W.S.: A Study in Attribution.* Cranbury, NJ: Associated University Presses.

Foster, D. (1996). "Response to Elliot and Valenza, "And Then There Were None". *Computers and the Humanities*, 30: 247-255.

Foster, D. (1999). The Claremont Shakespeare Authorship Clinic: How Severe Are the Problems. *Computers and the Humanities*, 32: 491-510.

Foster, D. (2000). *Author Unknown*. London: Macmillan.

Francis, I. S. (1966). And Exposition of a statistical approach to the *Federalist* Dispute. *The Computer and Literary Style.* (Leed, J. ed.). Kent State University Press.

Fries, C. C. (1952). *The Structure of English.* New York: Harcourt Brace.

Fucks, W. (1952). On Mathematical Analysis of Style. *Biometrika 39: 122-129.*

Fucks, W. (1954). On *Nahordnung* and *Fernordnung* in Samples of Literary Texts. *Biometrika,* 41:116-132.

Fucks, W. & Lauter, J. (1965). Mathematische Analyze des Literarischen Stils. In *Mathematik und Dichtung* (eds Kreuzer, H. & Gunzenhausers, R. Munich: Nymphenburger Verlagsbuckhandlung.

Good, I. J. (1965). Discussion of Morton 1965a. *Journal of the Royal Statistical Society A,* 128 :225-227.

Grant & Baker. (2001). Identifying reliable, valid markers of authorship: a response to Chaski. *Forensic Linguistics*, 8: 67-79.

Grieve, J. (2004). The Definition of a Language. *Proceedings of the 20th Northwest Linguistics Confrence:* 88-96.

Guiraud, H. (1954). *Les Caracteres Statistique du Vocabulaire.* Paris: Presses Universitaires de France.

Halliday, M.A.K. (1978). *Language as Social Semiotic: The Social Interpretation of Language and Meaning.* London: E. Arnold.

Hardcastle, R.A. (1993). Forensic Linguistics: An Assessment of the CUSUM Method for the Determination of Authorship. *Journal of the Forensic Science Society*, 33: 95-106.

Herdan, G. (1956). Chaucer's Authorship of the Equatorie of the Planetis: The Use of Romance Vocabulary as Evidence. *Language,* 32: 254-259.

Herdan, G. (1960). *Type Token Mathematics.* The Hague: Mouton & Co.

Herdan, G. (1965a). Discussion of Morton 1965a. *Journal of the Royal Statistical Society A,* 128:229-231.

Herdan, G. (1964). *Quantitative Linguistics.* London: Butterworth.

Herdan, G. (1966). *The Advanced Theory of Language as choice and Chance.* New York: Springer-Verlag.

Hilton, O. (1982). *Scientific Examination of Questioned Documents* (Revised Edition). New York: Elsevier: 1982.

Hilton, M. & Holmes, D. (1993). An Assessment of Cumulative Sum Charts for Authorship Attribution. *Literary and Linguistic Computing*, 8: 73-80.

Hockett, C. F. (1958). *A Course in Modern Linguistics.* New York: MacMillan.

Holmes, D. (1985). The Analysis of Literary Style—A Review. *The Journal of the Royal Statistical Society A,* 148: 328-341.

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

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

Holmes, D. (1998). The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing,* 13: 111-117.

Holmes, D. & Forsyth, R. (1995). The *Federalist* Revisited: New Directions in Authorship Attribution. *Literary and Linguistic Computing,* 10: 111-127.

Holmes, D. & Gordon, I. & Wilson, C. (2001). A Widow and her Soldier: Stylometry and the American Civil War. *Literary and Linguistic Computing,* 16: 403-420.

Holmes, D. & Robertson, M. & Paez, R. (2001). Stephen Crane and the New-York Tribune: A Case Study in Traditional and Non-Traditional Authorship Attribution. *Computers in the Humanities,* 35: 315-331.

Holmes, D. & Tweedie, F. (1995). Forensic Stylometry: A Review of the Cusum Controversy. In *Revue Informatique et Statistique dans les Science Humaine.* University of Liege, Belgium, pp. 19-47.

Honoré, A. (1979).Some Simple Measures of Richness of Vocabulary. *Association for Literary and Linguistic Computing Bulletin,* 7: 172-177.

Hoorn, J. F. et al. (1999). Neural Network Identification of Poets using Letter Sequences. *Literary and Linguistic Computing,* 14: 311-338.

Hoover, D. L. (2001). Statistical Stylistics and authorship Attribution: an Empirical Investigation. *Literary and Linguistic Computing,* 16: 421-443.

Hoover, D. L. (2002). Frequent Word Sequences and Statistical Stylistic. *Literary and Linguistic Computing,* 17:157-180.

Hoover, D. L. (2003). Frequent Collocations and Authorial Style. *Literary and Linguistic Computing,* 18: 261-286.

Horton, T.B. (1987). Doctoral Thesis. University of Edinburgh.

Huber, R.A (1956) *Crime Detection Laboratories: The Examination of Questioned Documents*, Seminar No. 4, Royal Canadian Mounted Police, Ottawa, 1956.

Hymes, D. (1974). *Foundations In Sociolinguistics.* Philadelphia: University of Pennsylvania Press.

Ingram, J. K. (1874). *Transactions of the New Shakespeare Society*, volume 1.

Johnson, R. (1979) Measures of Vocabulary Diversity. In *Advances in Computer-Aided Literary and Linguistic Research* (eds Ager, D. E. & Knowles, F. E. & Smith, J.) Birmingham: AMLC.

Johnson, P. F. (1974). The Use of Statistics in the Analysis of the Characteristics of Pauline Writings. *New Testament Studies,* 20: 92-100.

Keselj, V. & Peng, F. & Cercone, N. & Thomas, C. (2003). N-gram-based author profiles for authorship attribution. *Pacific Association for Computational Linguistics.*

Kenny, A. 1978.*The Aristotelian Ethics: A Study of the Relationship between the Eudemian and Nicomachean Ethics of Aristotle.* Oxford: Clarendon Press.

Kenny, A. (1982). *The Computation of Style.* Oxford: Pergamon Press.

Kenny, A. (1986). *A Stylometric Study of the New Testament.* Oxford University Press.

Khmelev, D. & Tweedie, F. J. (2001). Using Markov Chains for Identification of Writers, *Literary and Linguistic Computing,* 16: 299-308.

Kjell, B. (1994). Authorship Determination Using Letter-pair Frequency Features with Neural Network Classifiers. *Literary and Linguistic Computing,* 9: 119-124.

Kjetsaa, G. (1978). The Battle of *The Quiet Don*: Another Pilot Study. *Computers and the Humanities,* 11: 341-346.

Kjetsaa, G. (1979). *And Quiet Flows the Don* Through the Computer. *Association for Literary and Linguistic Computing Bulletin,* 1979: 248-256.

Koppel & Schler. (2003). Exploiting Stylistic Idiosyncrasies for Authorship Attribution. *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, Acapulco, Mexico.

Kreuzer, H. & Gunzenhausers (eds) (1965). *Mathematik und Dichtung.* Munich: Nymphenburger Verlagsbuckhandlung.

Kukushkina, Polikarpov & Khmelev (2002). Using Literal and Grammatical Statistics for Authorship Attribution. *Problems of Information Transmission:* 172-184.

Labov, W. (1994). *Principles of Linguistic Change.* Oxford: Blackwell.

Lake, D. (1975). *The Canon of Thomas Middleton's Plays.* London: Cambridge University Press.

Ledger, G. (1995). An Exploration of Differences in the Pauline Epistles using Multivariate Statistical Analysis. *Literary and Linguistic Computing,* 10: 85-97.

Ledger, G. & Merriam, T. (1994). Shakespeare, Fletcher, and the Two Noble Kinsmen. *Literary and Linguistic Computing,* 9: 119-124.

Leech, G. & Rayson, P. & Wilson, A. (2001). *Word Frequencies in Written and Spoken English.* London: Longman.

Leed, J. (ed) (1966). *The Computer and Literary Style.* Kent State University Press.

Lehmann, W. P. (1962) *Historical Linguistics: an Introduction.* New York: Holt, Rinehart and Winston.

Levison. M. & Morton, A. Q. & Wake, W.C. (1966) Some Statistical Features of the Pauline Epistles," *Journal of the Royal Philosophical Society* July,1966.

Levison, M. & Morton, A. Q. & Winspear, A. D. (1968). The Seventh Letter of Plato. *Mind,* 77: 309-325.

Lord, R.D. (1958). Studies in the History of Probability and Statistics VIII: De Morgan and the Statistical Study of Literary Style," *Biometrika* XLV:282.Love, H. (2002). *Attributing Authorship.* Cambridge University Press.

Lowe, D. & Matthew, R. (1995). A Stylometric Analysis by Radial Basis Functions. *Computers and the Humanities,* 29: 449-461.

Love, H. (2002). *Attributing Authorship: An Introduction.* Cambridge University Press.

M. (1890). Curves of Literary Style. *Science*, April 5, 1889: 269.

Malone, E. (1787) *A dissertation on parts one, two and three of Henry the Sixth tending to show that those plays were not written originally by Shakespeare.*

Manning, C. D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge, Mass. : MIT Press.

Martin, J. R. (2001). Language, Register and Genre. In Burns & Coffin (eds.) *Analysing English in a Global Context.* London: Routledge.

Mascol, C. (a.k.a. W. B. Smith). (1888a). Curves of Pauline and Pseudo-Pauline Style I. *Unitarian Review*, 30: 452-460.

Mascol, C. (a.k.a. W. B. Smith). (1888b). Curves of Pauline and Pseudo-Pauline Style II. Unitarian Review, 30: 539-546.

Matthews, R. & Merriam, T. (1993). Neural Computation in Stylometry I: An Application to the Works of Shakespeare and Fletcher. *Literary and Linguistic Computing* 8: 203-209.

Martindale, C. & McKenzie, D.P. (1995). On the Utility of Content Analysis in Authorship Attribution: The Federalist. *Computers and the Humanities,* 29:259-270.

McCarthy, C. (1985). *Blood Meridian.* London: Picador.

McColly, W. B. & Weier, D. (1983). Literary Attribution and Likelihood Ratio Tests: The Case of the Middle English Pearl-poems. *Computers and the Humanities,* 17:65-75.

McEnery, T. & Wilson, A. (1996). *Corpus Linguistics.* Edinburgh University Press.

McEnery, A. & Oakes, M. (2000). Authorship Studies/Textual Statistics. In *Handbook of Natural Language Processing* (eds Dale, R. & Moisl, H. & Somers, H.). Marcel Dekker.

McMenamin, G. (1993). *Forensic Stylistics.* Amsterdam: Elsevier Science Publishers.

McMenamin, G. (2002). *Forensic Linguistics.* Boca Raton, FL: CRC Press.

Mendenhall, T. C. (1887). The Characteristic Curves of Composition. *Science,* 11:237-249.

Mendenhall, T.C. (1901). A Mechanical Solution to a Literary Problem. *Popular Science Monthly,* 9:97-110.

Merriam, T. (1979). What Shakespeare Wrote in Henry VIII (Part I). *The Bard,* 2: 81-94.

Merriam, T. (1980). What Shakespeare Wrote in Henry VIII (Part II). *The Bard,* 2: 111-118.

Merriam, T. (1982). The Authorship of *Sir Thomas More. ALLC Bulletin,* 10: 1-7.

Merriam, T. (1986). The Authorship Controversy of Sir Thomas More: Smith on Morton. *Literary and Linguistic Computing,* 1: 104-106.

Merriam, T. (1987) Smith on Morton. *Literary and Linguistic Computing,* 1: 104-106.

Merriam, T. (1988). Was Hand B in *Sir Thomas More* Heywood's Autograph? *Notes and Queries,* 233: 455-458.

Merriam, T. (1989). An Experiment with the Federalist Papers. *Computing and the Humanities,* 23: 251-254.

Merriam, T. (1992). Doctoral Thesis. University of London.

Merriam, T. (1993). Marlowe's Hand in *Edward III. Literary and Linguistic Computing* 8: 59-72.

Merriam, T. (1994). Letter Frequency as a Discriminator of Authorship. *Notes and Queries,* 239: 467-469.

Merriam, T. (1996). Marlowe's Hand in *Edward III* revisited. *Literary and Linguistic Computing* 11: 19-22.

Merriam, T. (1997). Invalidation Reappraised. *Computers and the Humanities,* 30: 417-431.

Merriam, T. (1998). Heterogeneous authorship in Early Shakespeare and the problem of *Henry V. Literary and Linguistic Computing,* 13: 15-28.

Merriam , T. (2000). *Edward III. Literary and Linguistic Computing,* 15: 157-186.

Merriam, T. & Matthews, R. (1994). Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Marlowe. *Literary and Linguistic* Computing, 9: 1-6.

Michaelson, S. & Morton, A.Q. (1972a).The New Stylometry: A One-Word Test of Authorship for Greek Writers. *The Classical Quarterly,* 22: 89-102.

Michaelson, S. & Morton, A.Q. (1972b) The spans in between: A multiple test of authorship for Greek writers. *R.E.L.O. Review,* 1:23-77.

Michaelson, S & Morton, A.Q. (1972c). Last Words. *New Test Stud,* 18: pp. 192-208.

Michaelson, S. & Morton, A.Q. (1973). Positional Stylometry. in Aitken, A. J. & Bailey R. W. & Hamilton-Smith, N. (eds) (1973) *The Computer and Literary Studies*:69-83. Edinburgh University Press.

Michaelson, S. & Morton, A. Q. & Wake, W.C. (1978). Sentence length distributions in Greek hexameter and Homer. *ALLC Bulletin,* 6: 254.

Michéa, R. (1969). Repetition et variété dans l'emploi des mots. *Bulletin de la société de linguistique de Paris*, 61:1-21.

Michéa, R. (1971). De la relation entre le nombre des mots d'une fréquence determinée et celui des mots differents employés dans le texte. *Cahiers de Lexicologie,* 18: 65-78.

Miller, C. & Swift, K. (1971). De-Sexing the English Language. *New York Magazine.*

Monsarrat, G. (2002). A Funeral Elegy: Ford, W.S., and Shakespeare. *Review of English Studies,* 53: 186.

Moritz, R. E. (1903). On the Variation and Functional Relationship of Certain Sentence-Constants in Standard Literature. *University Bulletin,* 8: 229-253.

Morton, A. Q. (1965a). The Authorship of Greek Prose. *Journal of the Royal Statistical Society A,* 128:169-233.

Morton, A. Q. (1965b) The Integrity of the Pauline Epistles. *Manchester Statistical Society,* March 1965.

Morton, A. Q. (1965c). *The Authorship of the Pauline Epistles: A Scientific Solution.* Saskatoon: University of Saskatchewan.

Morton, A. Q. (1978). *Literary Detection: How to Prove Authorship and Fraud in Literature and Documents*. New York: Scribners.

Morton, A. Q. (1986). Once. A Test of Authorship Based on Words which are not Repeated in the Sample. *Journal of the Association for Literary and Linguistic Computing,* 1: 1-8.

Morton, A. Q. (1991). Proper Words in Proper Places. Technical Report 91/r18, University of Glasgow, Computing Science Department.

Morton, A.Q. & Levison, M. (1966) "Some Indicators of Authorship in Greek Prose. In Leed (ed). *The Computer and Literary Style.* Kent State University Press.141-179.

Morton, A. Q. & McLeman, J. (1964). *Christianity in the Computer Age.* New York: Harper & Row, Publishers.

Morton, A. Q. & McLeman, J. (1966). *Paul, The Man and the Myth.* New York: Harper and Row.

Morton, A.Q. & Winspear, A. & Levison, M. & Michaelson, S. (1971). *It's Greek to the Computer*. Montreal: Harvest House.

Morton, A.Q. & Michaelson, S. (1990). The Qsum Plot. Technical Report CSR-3-90, University of Edinburgh.

Morton, A. Q. & Michaelson, S. & Hamilton-Smith (1977). To Couple is the Custom. Technical Report CSR-22-79. University of Edinburgh.

Mosteller, F. & Wallace, D. (1963). Inference in an Authorship Problem. *Journal of The American Statistical Association,* 58: 275-309.

Mosteller, F. & Wallace, D. (1964). *Inference and Disputed Authorship: The Federalist* (1st Edition). Reading, MA: Addison-Wesley.

Mosteller, F. & Wallace, D. (1984). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers* (2nd Edition). New York: Springer-Verlag.

Neumann, K. J. (1990). *The Authenticity of the Pauline Epistles in the Light of Stylostatistical Analysis.* Atlanta: SBL, Scholar's Press.

Niederkorn (2002). A Scholar Recants on His 'Shakespeare' Discovery. *New York Times,* 20 June 2002: B1 and B5.

Oakes, P. (1998). *Statistic for Corpus Linguistics.* Edinburgh University Press.

Oakman, R. (1980). *Computer Methods for Literary Research.* Columbia: University of South Carolina Press.

O'Brien, D. P. & Darnell, A. C. (1982). *Authorship Puzzles in the History of Economics: a Statistical Approach.* London: Macmillan.

O'Donnell, B. (1966). Stephen Crane's *The O'Ruddy*: a problem in authorship discrimination. In Leed (ed.) *The Computer and Literary Style.* Kent State University Press: 107-115.

Osborn, A. S. (1910) *Questioned Documents* (First edition). Rochester, NY: Lawyers Cooperative Publishing.

Osborn, A. S. (1929). *Questioned Documents.* (Second edition). Albany, NY: Biyd Printing Co.

Parker, H.A. (1890). Curves of Literary Style. *Science,* 13 (321): 245.

Peng, F. & Schuurmans, D. & Keselj, V. & Wang, S. (2003). Language independent authorship attribution using character level language models. *Tenth Conference of the European Chapter of the Association for Computational Linguistics.*

Petyt, K. M. (1980). The Study of Dialect: An Introduction to Dialectology. London: A. Deutsch.

Pollatschek, M. & Radday, Y. T. (1981). Vocabulary Richness and Concentration in Hebrew Biblical Literature. *Association for Literary and Linguistic Computing Bulletin,* 8: 217-231.

Pollatschek & Radday, Y. T. 1985. Vocabulary Richness and Concentration. In *Genesis – an Authorship Study* (eds Radday, Y. T. & Shore, H.). Rome: Biblical Institute.

Radday, Y. T. (1970). Isaiah and the Computer: A Preliminary Report. *Computers and the Humanities,* 5: 65-73.

Radday, Y. T. & Shore, H. (eds) (1985) *Genesis – an Authorship Study.* Rome: Biblical Institute.

Robins, R. H. (1964). *General Linguistics: An Introductory Survey.* London: Longmans.

Roderick, Richard. (1758). In Edwards, T. *Canons of Criticism:* 225-228.

Rudman, J. (1998). The State of Authorship Attribution Studies: Some Problems and Solutions. *Computers and the Humanities,* 31: 351-365.

Rudman, J. (1998). Non-Traditional Authorship Attribution Studies in the 'Historia Augusta': Some Caveats. *Literary and Linguistic Computing,* 13: 151-157.

Sapir, E. (1949). *Language.* New York: Harcourt, Brace and Company.

Sams, E. (1994). *Edmund Ironside* and Stylometry. *Notes and Queries,* 239: 469-470.

Sandord, A.J. & Aked, J.F. & Moxley, L.M. & Mullin, J. (1994). A Critical Examination of Assumptions Underlying the Cusum Technique of Forensic Linguistics. *Forensic Linguistics,* 1: 151-167.

Schachter, P. (1985). Parts-of-speech systems. In Shopen, T (ed.) *Language Typology and Syntactic Description: Clause Structure:* 3-61.

Sherman, L. A. (1893). *Analytics of Literature.* Boston: Ginn.

Shopen, T (ed.) *Language Typology and Syntactic Description: Clause Structure.* Cambridge University Press.

Sichel, H.S. (1975). On a Distribution Law for Word Frequencies. *Journal of the American Statistical Association,* 70: 542-547.

Siegel, S. (1956). *Nonparametric Statistics.* New York: McGraw-Hill.

Simpson, E. H. (1949). Measurement of Diversity. *Nature,* 163: 688.

Smith, M. W. A. (1983). Recent Experience and New Developments of Methods for the Determination of Authorship. *Association for Literary and Linguistic Computing Bulletin,* 11: 73-82.

Smith, M.W.A. (1984). Critical Reflections on the Determination of Authorship by Statistics. *The Shakespeare Newsletter,* 34: 4, 5, 28, 33, 34, 47.

Smith, M.W.A. (1985a). An Investigation of the Basis of Morton's Method for the Determination of Authorship. *Style,* 19: 341-368.

Smith, M.W.A. (1985b). An Investigation of Morton's Method to Distinguish Elizabethan Playwrights. *Computers and the Humanities,* 19: 3-21.

Smith, M.W.A. (1987a). *Hapax Legomena* in Prescribed Positions: an Investigation of Recent Proposals to resolve problems of authorship. *Journal of the Association for Literary and Linguistic Computing,* 2: 145-152.

Smith, M.W.A. (1987b). The Authorship of *Pericles:* New Evidence for Wilkins. *Journal of the Association for Literary and Linguistic Computing,* 2: 221-230.

Smith, M.W.A. (1987c). Merriam's Application of Morton's Method. *Computers and the Humanities,* 21: 59-60.

Smith, M.W.A. (1987d). *The Revenger's Tragedy:* The Derivation of Methods for the Determination of Authorship. *Computers and the Humanities,* 21: 21-55.

Smith, M.W.A. 1988. The Authorship of Acts I and II of *Pericles*: a New Approach Using First Words of Speeches. Computers and the Humanities, 22: 1.

Smith, M.W.A. (1989a). A Procedure to Determine Authorship Using Pairs of Consecutive Words: more Evidence for Wilkins' Participation in *Pericles. Computers and the Humanities,* 23: 113-129.

Smith, M.W.A. (1990). Attribution by Statistics: A Critique of Four Recent Studies. *Revue Informatique et Statistique dans les Science Humaine,* 26: 233-251.

Smith, M.W.A. (1991). The Authorship of *The Revenger's Tragedy. Notes and Queries,* 236:508-511.

Smith, M.W.A. (1992). The Problem of Acts I-II of *Pericles. Notes and Queries,* 237: 346-355.

Smith, M.W.A. (1993). *Edmund Ironside. Notes and Queries,* 238: 202-205.

Smith, M.W.A. (1994). *Sir Thomas More, Pericles* and Stylometry, *Notes and Queries,* 239: 55-58.

Smith, W. B. (1888a). See Mascol (1888a).

Smith, W. B. (1888b). See Mascol (1888b).

Somers, H. & Tweedie, F. (2003). Authorship Attribution and Pastiche. *Computers and the Humanities*, 37: 407-429.

Spedding, J. (1850). Who Wrote Shakespeare's *Henry VIII. The Gentleman's Magazine,* 115-123.

Stamatatos, E. & Fakotakis, N. & Kokkinakis, G. (2000). Automatic Text Categorization in Terms of Genre and Author. *Computational Linguistics,* 26: 471-495.

Stamatatos, E. & Fakotakis, N. & Kokkinakis, G. (2001). Computer-Based Authorship Attribution Without Lexical Measures. *Computers and the Humanities,* 35: 193-214.

Svartvik, J. (1968). *The Evans Statement.* Gothenburg: University Gothenburg.

Swinburne, Algernon. Report of the Proceedings on the First Anniversary Session of the Newest Shakespeare Society. 1876.

Tallentire, D. R. (1973). Towards and Archive of Lexical Norms—A Proposal. In *The Computer and Literary Studies* (eds Aitken, A. J. & Bailey R. W. & Hamilton-Smith, N.). Edinburgh University Press.

Tarlinskaja, M. (1987). *Shakespeare's Verse: Iambic Pentameter and the Poet's Idiosyncrasies.* New York: Peter Lang.

Thompson, Hunter S. (1973). *Fear and Loathing: On the Campaign Trail '72.* New York: Popular Library.

Thorndike, Ashley H. (1901). *The Influence of Beaumont and Fletcher on Shakespeare.* New York: AMS Press. (1966 Edition).

Totty, R.N. & Hardcastle, R.A. & Pearson, J. (1987). Forensic Linguistics: the determination of authorship from habits of style. *Journal of the Forensic Society,* 27: 13.

Tuldava, J. (1977). Quantitative Relations between the Size of the Text and the Size of the Vocabulary. *SMIL Quarterly, Journal of Linguistic Calculus*, 4.

Tweedie, F. & Baayen, H. (1998). How Variable may a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities,* 32:323-53.

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

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

Tweedie, F. J. & Singh, S. & Holmes, D. I. (1996b). An Introduction to Neural Networks in Stylometry. *Research in Humanities Computing,* 5: 249-269.

Usher, S. & Najock, D. (1982).A Statistical Study of Authorship in the Corpus Lysiacum. *Computers and the Humanities*, 16: 85-105.

Vickers, B. (2002). *Counterfeiting Shakespeare.* Cambridge University Press.

Wake, W. C. (1957). Sentence-length Distributions of Greek authors. *Journal of the Royal Statistical Society A,* 120: 331-346.

Waugh, S. & Adams, A. & Tweedie, F. (2000). Computational stylistics Using Artificial Neural Networks. *Literary and Linguistic computing,* 15: 187-198.

Weber, H. (1812). *The Works of Beaumont and Fletcher.*

Whissell, C. (1996). Traditional and Emotional Stylometric Analysis of the Songs of Beatles Paul McCartney and John Lennon. *Computers and the Humanities,* 30: 257-265.

Whitney, W. D. (1901). *Language and the Study of Language.* New York: AMS Press.

Williams, C.B. (1940).A Note on the Statistical Analysis of Sentence-length as a Criterion of Literary Style. *Biometrika,* 31:363-390.

Williams, C.B. (1970). *Style and Vocabulary.* New York: Hafner Publishing Co.

Woods, A. & Fletcher, P. & Hughes, A. (1986). *Statistics in Language Studies.* New York: Cambridge University Press.

Yule, G. U. (1939). On Sentence-length as a Statistical Characteristic of Style in Prose, with Application to Two Cases of Disputed Authorship. *Biometrika,* 31: 356-361.

Yule, G. U. (1944). *The Statistical Study of Literary Vocabulary.* Cambridge University Press.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort.* Cambridge, MA: Addison-Wesley Publishing Co.

APPENDIX

Each variant of each type of textual measurement that was tested in this study is listed in this Appendix. Each entry consists of the following information: a code number, which corresponds to the code number found in the accuracy tables provided in Chapter 5; a description of the measurement; and, in brackets, the number of individual measurements included in that set of textual measurements. When the variant is based on the values of multiple textual measurements, a list of the strings of characters whose relative frequencies are being measured is also included in the entry, except in the case of the n-gram variants, which are based on the frequency of far too many strings to be listed here. In addition, when available, references are included to a selection of attribution studies that have made use of that set of textual measurement in the past.

1 Average Word Length (1)
(De Morgan 1851, Foster 1989)

2 Average Sentence Length in Words (1)
(Eddy 1888, Sherman 1888, 1893)

3 Average Sentence Length in Characters (1)
(Smith 1888)

4 Word-Length (Distribution) Profile (15)
(Mendenhall 1887, 1901, Brinegar 1963, Williams 1970)
1-, 2-, 3-, 4-, 5-, 6-, 7-, 8-, 9-, 10-, 11-, 12-, 13-, 14-, 15-character words.

5 Word-Length (Distribution) Profile (10)
1-, 2-, 3-, 4-, 5-, 6-, 7-, 8-, 9-, 10-character words.

6 Word-Length (Distribution) Profile (5)
1-, 2-, 3-, 4-, 5-character words.

7 Sentence-Length (Distribution) Profile in Words (10)
(Yule 1939, Williams 1940, Wake 1957, Morton 1965)
1- to 5-, 6- to 10-, 11- to 15-, 16- to 20-, 21- to 25-, 26- to 30-, 31- to 35-,36- to 40-, 41- to 45-, 46- to 50-word sentences.

8 Sentence-Length (Distribution) Profile in Words (6)
1- to 5-, 6- to 10-, 11- to 15-, 16- to 20-, 21- to 25-, 26- to 30-word sentences.

9 Sentence-Length (Distribution) Profile in Words (5)
1- to 10-, 11- to 20-, 21- to 30-, 31- to 40-, 41- to 50-word sentences.

10 Sentence-Length (Distribution) Profile in Words (3)
1- to 10-, 11- to 20-, 21- to 30-word sentences.

11 Sentence-Length (Distribution) Profile in Characters (12)
1- to 25-, 26- to 50-, 51- to 75-, 76- to 100-, 101- to 125-, 126- to 150-, 151- to 175-, 176- to 200-, 201- to 225-, 226- to 250-, 251- to 275-, 276- to 300-character sentences.

12 Sentence-Length (Distribution) Profile in Characters (8)
1- to 25-, 26- to 50-, 51- to 75-, 76- to 100-, 101- to 125-, 126- to 150-, 151- to 175-, 176- to 200-character sentences.

13 Sentence-Length (Distribution) Profile in Characters (6)
1- to 50-, 51- to 100-, 101- to 150-, 151- to 200-, 201- to 250-, 251- to 300-character sentences.

14 Sentence-Length (Distribution) Profile in Characters (4)
1- to 50-, 51- to 100-, 101- to 150-, 151- to 200-character sentences.

15 Unrestricted Type-Token Ratio (1)

16 Restricted Type-Token Ratio (1) *First 119 Words.*

17 Yule's K and Simpson's D (1)
(Yule 1944, Simpson 1949, Holmes 1992)

18 Guiraud's R (1)
(Guiraud 1954, Holmes 1992)

19 Herdan's C (1)
(Herdan 1960, 1965)

20 Dugast's k (1)
(Guiraud 1954)

21 Honoré's H (1)
(Honoré 1979)

22 Sichel's S and Michéa's M (1)
(Sichel 1975, Michéa 1969, 1971, Holmes 1992)

23 Entropy (1)

24 Tuldava's LN (1)
(Tuldava 1977)

25 W (1) $a = -0.165$
(Brunet 1978)

26 W (1) $a = -0.168$

27 W (1) $a = -0.172$

28 Grapheme Profile (26)
(Yule 1944, Herdan 1966, Merriam 1988, 1998, Merriam & Ledger 1994)
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

29 Single-Position Grapheme Profile: First Grapheme in Word (25)
(Yule 1944, Herdan 1966)
A B C D E F G H I J K L M N O P Q R S T U V W Y Z

30 Single-Position Grapheme Profile: Second Grapheme in Word (24)
A B C D E F G H I K L M N O P Q R S T U V W X Y

31 Single-Position Grapheme Profile: Third Grapheme in Word (26)
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

32 Single-Position Grapheme Profile: Last Grapheme in Word (22)
(Ledger 1995)
A B C D E F G H I K L M N O P R S T U W X Y

33 Single-Position Grapheme Profile: Second to Last Grapheme in Word (24)
A B C D E F G H I K L M N O P R S T U V W X Y Z

34 Single-Position Grapheme Profile: Third to Last Grapheme in Word (25)
A B C D E F G H I J K L M N O P R S T U V W X Y Z

35 Multi-Position Grapheme Profile: First Three Graphemes in Word (66)
A B C D E F G H I K L M N O P R S T U V W Y in the first, second and third word-positions.

36 Multi-Position Grapheme Profile: First Six Graphemes in Word (132)
A B C D E F G H I K L M N O P R S T U V W Y in the first, second, third, fourth, fifth and sixth word-positions.

37 Multi-Position Grapheme Profile: Last Three Graphemes in Word (66)
A B C D E F G H I K L M N O P R S T U W X Y in the last, second to last and third to last word-positions.

38 Multi-Position Grapheme Profile: Last Six Graphemes in Word (132)
A B C D E F G H I K L M N O P R S T U W X Y in the last, second to last, third to last, fourth to last, fifth to last and sixth to last word-positions.

39 Multi-Position Grapheme Profile: First and Last Six Graphemes in Word (264)
A B C D E F G H I K L M N O P R S T U V W Y in the first, second, third, fourth, fifth and sixth word-positions; and A B C D E F G H I K L M N O P R S T U W X Y in the last, second to last, third to last, fourth to last, fifth to last and sixth to last word-positions.

40 Word Internal Grapheme Profile (26)
(Ledger 1995)
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

41 (Two-Limit) Word Profile (265)
(Smith 1888, Ellegård 1962, Mosteller & Wallace 1964, Morton 1965, Burrows 1988)
a i s t children same does work says just done really family down also day would second public three might did very about being home school say better see days 10 she else before half hand long hard look have where lost called find thought friends end rather country could keep the almost last too their because far something two every am an as at few be by there these do go he if in is it asked told for given me mr took my past no of on or so to up us national use take seen get another britain into got right other become week well went under had h as were first her anything him his such young was way how place head years who why over sure which human 000 government sense again more house service while most great made give make left people start after until many state wrong its true less through middle thing think little making enough job yet what order without when both back themselves should four matter will comes with let high man may men london never nothing life during local like than that point them then they this used free new taken even ever seems from moment come away not now much between know been call came time must though off good case year once old best words one only against british world our out feel own probably all and any are those bad put going among need quite street some but full still course having mind news doing can next times least said

42 Five-Limit Word Profile (144)
up a i new s what even without take when from does come not now another back just much into between right should know down also other been week day would time must three might good well did year once about under being home had has old were first say best one see only will him his british before such world out own was way how with years who all over and any long which are have where those thought more house end while most great made could the put may make people last after too many their its never because far two every an as at few be by less there these do some but like go he still through if than in is that it think for point them then they no of on or this can next so least to said

43 Ten-Limit Word Profile (85)
up who all a i new s over and any what which are when have from those not now more most much made could into the other people been after too would time their its because well two about an as at being be by had there has do were first say some but like one he if only in that is it will him his for such them then out they no of on or was way how this with can so to said

44 Fifteen-Limit Word Profile (60)
up would who all time a their s and any what which are about when have an as at from be by had there has do were not but like more one he much if only in is that it will his for them then out they no the of on or was this with can so to people been

45 Twenty-Limit Word Profile (45)
would he who all if a in their is that it s and his for out they no are the about of on or when have an as at was from be this by there with has can so not to but more one been

46 Twenty-Five-Limit Word Profile (36)
are the of on he have who an as at if a was in from their is that it be s this by there and with has his for so not to but more one they

47 Thirty-Limit Word Profile (23)
are the of on have an as at a was in that is it be s by and with for to but one

48 Forty-Limit Word Profile (5)
and the to a of

49 Punctuation Mark Profile by Punctuation Marks (8)
(Smith 1888, O'Donnell 1966)
. , ; : - ‘ (?

50 Punctuation Mark Profile by Words (8)
. , ; : - ‘ (?

51 Punctuation Mark Profile by Characters (8)
. , ; : - ‘ (?

52 Grapheme and Punctuation Mark Profile (34)
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z . , ; : - ‘ (?

53 Five-Limit Word and Punctuation Mark Profile (152)
Profile 38 and Profile 42.

54 Ten-Limit Word and Punctuation Mark Profile (93)
Profile 38 and Profile 43.

55 Twenty-Limit Word and Punctuation Mark Profile (53)
Profile 38 and Profile 45.

56 Single-Position Word Profile: First Word in Sentence (19)
(Morton 1965, 1978)
A I There As At Not But One He If In It For They The No This So To

57 Single-Position Word Profile: Second Word in Sentence (21)
I a s are have the there one he if in that is it for they of was this can to

58 Single-Position Word Profile: Third Word in Sentence (22)
a s and are have not the as at be has he in that is it for they of was this to

59 Single-Position Word Profile: Fourth Word in Sentence (22)
a s and are have not the been be has he in is that it for of on was this with to

60 Single-Position Word Profile: Last Word in Sentence (2)
(Morton & Michaelson 1972a)
it them

61 Single-Position Word Profile: Second to Last Word in Sentence (20)
a s and are the their as at be in is it his for of on or was with to

62 Single-Position Word Profile: Third to Last Word in Sentence (22)
a s and are have from the their an at by he in is it his for of on was with to

63 Single-Position Word Profile: Fourth to Last Word in Sentence (22)
a s and have not the about as at be by in that is it his for of on was with to

64 Multi-Position Word Profile: First Four Words in Sentence (85)
0There 3with 1the 3have 3the 0The 1for 1he 1if 1in 1is 1it 3for 1of 0A 0I 1I 1to 1a 1s 2a 0For 2s 3a 3that 3s 3this 2as 2at 2be 3not 2he 2in 2is 2it 0Not 2no 2of 2to 1one 3has 3been 1was 3be 3he 3in 3is 3it 0One 3was 1are 3of 3on 1have 3to 3and 3are 2the 0They 0This 1that 2for 1can 1they 1this 0But 2not 2have 1there 0As 0At 2has 0He 0If 0In 0It 0No 2was 2that 0So 0To 2they 2and 2this 2are

65 Multi-Position Word Profile: First Eight Words in Sentence (168)

0There 7of 7on 1the 7to 6and 3have 6are 5the 1for 7with 0A 0I 1I 1a 1s 2a 2s 3a 3s 5for 4a 4s 5a 5s 6a 6s 7a 7s 3not 0Not 7that 1was 0One 3and 5was 3are 4that 2the 7been 7and 0They 0This 6the 5have 1that 2for 1they 1this 0But 6for 2have 4not 0As 0At 2has 6with 0He 0If 0In 0It 0No 2was 0So 0To 6has 3with 4and 6was 4are 3the 6that 0The 1he 7the 1if 1in 1is 1it 3for 1of 1to 0For 7have 3that 3this 2as 2at 2be 7for 2he 2in 2is 2it 2no 2of 2to 4have 1one 3has 3been 3be 3he 3in 3is 3it 3was 1are 3of 3on 1have 3to 7his 5and 4as 4at 7was 4be 4by 4he 7who 4in 4is 4it 5with 4the 4of 4on 4to 1can 5at 5be 5by 4for 5he 5in 5is 5it 5of 5on 2not 5to 5that 6as 6at 6be 6by 1there 6in 6is 6it 6of 6on 6to 4has 6have 2that 2they 7at 7be 7by 2and 2this 4was 2are 7in 7is 7it

66 Multi-Position Word Profile: Last Four Words in Sentence (66)

3with 1the 2from 0it 3have 3the 1for 1as 1at 1be 1in 1is 1it 3for 1of 1on 1or 1to 1a 1s 2a 2s 3a 3that 3s 2an 2at 3not 2by 2he 2in 2is 2it 2of 2on 2to 2their 0them 1his 1was 3as 3at 3be 3by 3his 1with 1and 3in 3is 3it 3was 1are 3of 3on 3to 3and
2the 2for 3about 2with 2have 1their 2his 2was 2and 2are

67 Multi-Position Word Profile: Last Eight Words in Sentence (157)

7of 7on 1the 2from 7to 6and 3have 6are 5the 1for 7with 1a 1s 2a 2s 3a 3s 5for 4a 4s 5a 5s 6a 6s 7a 3not 4with 2their 1his 7that 7they 1was 1with 5his 3and 5was 4that 4they 2the 7and 7are 6the 2for 3about 6for 2have 4not 6with 2his 2was 6has 3with 4and 6was 0it 3the 6that 1as 1at 1be 7the 5more 1in 1is 1it 3for 6from 1of 1on 1or 1to 7have 3that 6been 2an 2at 7for 2by 2he 2in 2is 2it 5not 2of 2on 2to 4have 0them 3as 3at 3be 3by 3his 1and 3in 3is 3it 3was 1are 3of 3on 3to 7his 5and 4as 7was 4be 4by 5are 7who 4in 4is 4it 5with 4the 4of 4on 4to 5as 5at 5be 4for 5in 5is 5it 2with 5of 5on 5to 1their 5that 6as 5they 6be 6by 6he 6in 6is 6it 6of 6on 6to 4has 6have 4his 7as 7at 7be 7by 2and 4was 6one 2are 7he 7in 7is 7it

68 Two-Word Collocation Profile (102)

(Morton 1978, O'Brien & Darnell 1982, Smith 1989, Hoover 2002)

and_the by_the the_same a_few by_a with_the to_be to_a not_to to_do seems_to in_this and_a was_the as_a and_that was_a at_a to_take of_a on_the not_the had_to have_a in_their more_than is_that to_make of_them but_it be_a out_of there_was is_not the_other there_is a_new there_are in_the for_a the_first he_would on_to of_the this_is the_world would_be they_were in_an to_say it_was have_to to_see he_had he_was it_is he_has that_is has_been will_be that_it to_the that_they should_be all_the about_the from_a into_the when_he would_have on_a is_the if_the in_a as_the up_to if_it the_right it_would at_least the_new to_have that_a is_a the_only for_the over_the at_the in_his but_the of_his with_a when_the from_the to_get that_the as_they to_his one_of the_time have_been he_is

69 Three-Word Collocation Profile (1)

one_of_the

70 Two-Limit Two-Gram Profile (469)

(Bennett 1976, Clement & Sharp 2003, Keselj *et al.* 2003)

71 Ten-Limit Two-Gram Profile (350)

72 Twenty-Limit Two-Gram Profile (298)

73 Two-Limit Three-Gram Profile (2051)

74 Ten-Limit Three-Gram Profile (1081)