THE UNIVERSITY *of* EDINBURGH

The Effectiveness of the Stylometry of Function Words in Discriminating between Shakespeare and Fletcher

Thomas Bolton Horton

Ph D
University of Edinburgh
1987

Abstract

A number of recent successful authorship studies have relied on a statistical analysis of language features based on function words. However, stylometry has not been extensively applied to Elizabethan and Jacobean dramatic questions. To determine the effectiveness of such an approach in this field, language features are studied in twenty-four plays by Shakespeare and eight by Fletcher. The goal is to develop procedures that might be used to determine the authorship of individual scenes in *The Two Noble Kinsmen* and *Henry VIII*.

Homonyms, spelling variants and contracted forms in old-spelling dramatic texts present problems for a computer analysis. A program that uses a system of pre-edit codes and replacement/expansion lists was developed to prepare versions of the texts in which all forms of common words can be recognized automatically.

To evaluate some procedures for determining authorship developed by A. Q. Morton and his colleagues, occurrences of 30 common collocations and 5 proportional pairs are analyzed in the texts. Within-author variation for these features is greater than had been found in previous studies. Univariate chi-square tests are shown to be of limited usefulness because of the statistical distribution of these textual features and correlation between pairs of features. The best of the collocations do not discriminate as well as most of the individual words from which they are composed.

Turning to the rate of occurrence of individual words and groups of words, distinctiveness ratios and t-tests are used to select variables that best discriminate between Shakespeare and Fletcher. Variation due to date of composition and genre within the Shakespeare texts is examined. A multivariate and distribution-free discriminant analysis procedure (using kernel estimation) is introduced. The classifiers based on the best marker words and the kernel method are not greatly affected by characterization and perform well for samples as short as 500 words. When the final procedure is used to assign the 459 scenes of known authorship (containing at least 500 words), almost 95% are assigned to the correct author. Only two scenes are incorrectly classified, and 4.8% of the scenes cannot be assigned to either author by the procedure.

When applied to individual scenes of at least 500 words in *The Two Noble Kinsmen* and *Henry VIII*, the procedure indicates that both plays are collaborations and generally supports the usual division. However, the marker words in a number of scenes often attributed to Fletcher are very much closer to Shakespeare's pattern of use. These scenes include *TNK* IV.iii and *H8* I.iii, IV.i–ii and V.iv.

To the memory of my mother, who always believed in the members of her family, even when we ourselves doubted.

Preface

A few initial words about this document may clarify some issues for the reader. First, although I have been engaged in a course of study at a British university, I have followed my native American spelling and punctuation conventions in writing this dissertation. This decision was taken solely for reasons of consistency. My supervisor and I recognized that it would be easier for me to be consistently American than British (although this has not proved as easy I imagined).

Second, this dissertation was prepared using the LaTeX and TeX document preparation software systems, and was printed on a Canon desktop laserprinter (driven by software developed in the Computer Science department). All tables and graphs in the document were created using LaTeX, with the exception of Figure 6–2 and the curved line in Figure 6–1. For the most part LaTeX has proved very satisfactory. However, I discovered rather late that the associated BibTeX program, which takes care of the citations in the text and the bibliography, was not as sophisticated as I had hoped.

Citations in the text are composed of bracketed numbers, which correspond to works listed in the numbered bibliography at the end of the volume. Optionally, the citation is followed by a page number (or numbers). For example, I might use this method to refer to the section in Mary-Claire van Leunen's *A Handbook for Scholars* that discusses why I should not hesitate to use first-person pronouns for the sake of clarity [164, pp. 37-41]. (This book contains some excellent advice for improving scholarly writing, which I have attempted to take to heart.) The bibliography is arranged alphabetically by author, with the reference numbers printed before each entry. For editions of plays, an unnumbered cross-reference to the text is listed under the editor's surname.

Acknowledgements

I've always admired short and witty acknowledgments, but looking back over almost five years of research, I find that I am indebted to far too many people to be brief. At the risk of this becoming "Chapter 9" I'll begin....

My interest in this topic dates from my undergraduate days at the University of Tennessee, where Dr. Norman Sanders and Dr. Chuck Pfleeger encouraged me in my interdisciplinary interests. Many thanks to them and everyone else associated with the College Scholars program for helping to plant the seeds.

Without the funding of the Marshall Aid Commemoration Commission, I would have never come to Britain to study. In particular, Geraldine Cully has given me support and friendship that I value greatly.

Without machine-readable texts of Shakespeare and Fletcher, this dissertation would have been somewhat more brief. Many thanks to the Oxford University Shakespeare Department for allowing me to access their Shakespeare texts; in addition, Gary Taylor provided me with very useful advice on textual matters in the very early stages. I am also indebted to the Text Archive at the Oxford University Computing Service, especially Lou Burnard, for supplying me with copies of these and other texts. In addition, Prof. David Gunby of the University of Canterbury (New Zealand) has generously provided copies of a number of machine-readable dramatic texts that were created as part of the preparation for the Cambridge edition of Webster's works.

The academic environment in which I have pursued my studies has been excellent, primarily because of a number of people in the Computer Science department and the Edinburgh Regional Computing Centre. Many thanks to George Cleland for facilitating access to departmental machines and software, and to Neil Hamilton-Smith for a great deal of assistance with the CONCORD concordance program. The Computer Science department's generosity in funding travel expenses for a number of conferences will not be forgotten.

Not only has Chris Robinson been a tremendous friend, her assistance with all matters concerning English Language has been invaluable. Her careful reading of

most of the draft of this dissertation has rescued me from a number of mistakes. Others who have made useful comments on sections of the draft are Andrew Morton and Tom Merriam (who also supplied me with copies of a number of articles which I could not access).

Through the years Nik Traub has been an excellent friend and a valuable source of advice on computing. Without his advice and encouragement regarding Unix, I would have probably never developed the skills that allowed me to successfully complete this study. He also demonstrated that PhD students *can* actually finish, and I'll treasure his guacamole recipe forever.

A number of other friends have significantly contributed to my well-being. In the early days, Stacy Waters and Susan Aronstein provided academic stimulation and welcome distraction (liquid and culinary, respectively). Since then there have been a number of others who have worked hard to keep me sane, including Graeme Steele, Eva Ashford (meow), Eric Wilson (woof), Mark Davoren, Laurent Langlois and Lise Desjardins. Eric deserves special notice for showing me glimpses of Scotland that I'd probably have missed, including a number of memorable pubs from Lancaster to Leith.

Dr. Colin Aitken of the Statistics Department has been outstandingly patient and helpful in advising me about discriminant analysis. (Not to mention the fact that he let me use his program.)

My supervisor, Prof. Sidney Michaelson, has made me feel welcome and at ease since the day I arrived at the university. His remarkable breadth of knowledge has been a great encouragement to me, and I appreciate the hours that he has spent in developing software for me and reading my work.

Andrew and Jean Morton have been exceptionally kind and generous, and some of my best memories of Scotland will be of times spent at the Abbey Manse. I'll try to remember Andrew's curiosity and energy whenever I need inspiration in my work.

Many thanks to my family, who have always communicated their pride and support.

Finally, my appreciation and admiration for Viola has grown with each year of separation and increasing phone bills. The combination of love and understanding she has shown me has been a most precious gift, and I certainly wouldn't have made it through this experience intact without it.

Table of Contents

List of Abbreviations of Titles

Shakespeare:

1H4	*The First Part of King Henry the Fourth*
Ant	*Antony and Cleopatra*
AWW	*All's Well that Ends Well*
AYL	*As You Like It*
CE	*The Comedy of Errors*
Cor	*Coriolanus*
Cym	*Cymbeline*
H5	*The Life of King Henry the Fifth*
JC	*Julius Caesar*
KJ	*The Life and Death of King John*
LLL	*Love's Labor's Lost*
Mac	*Macbeth*
MAN	*Much Ado about Nothing*
MND	*A Midsummer Night's Dream*
MV	*The Merchant of Venice*
MWW	*The Merry Wives of Windsor*
R2	*The Tragedy of King Richard the Second*
R3	*The Tragedy of King Richard the Third*
Rom	*Romeo and Juliet*
Tem	*The Tempest*
TGV	*The Two Gentlemen of Verona*
TN	*Twelfth Night*
TS	*The Taming of the Shrew*
WT	*The Winter's Tale*

Fletcher:

Bond	*Bonduca*
Chan	*The Chances*
Deme	*Demetrius and Enanthe*
Prin	*The Island Princess*
Priz	*The Woman's Prize*
Subj	*The Loyal Subject*
Thom	*Monsieur Thomas*
Vale	*The Tragedy of Valentinian*

Disputed:

H8	*The Famous History of the Life of King Henry the Eighth*
TNK	*The Two Noble Kinsmen*

List of Tables

List of Figures

Chapter 1

Introduction

The purpose of this dissertation is to explore the possibilities of using a scientific and statistical approach to solve a Shakespearean authorship question. There are many unanswered authorship questions involving Renaissance English dramatic texts, and the answers to these questions would obviously be very valuable to the literary critic and the theatrical historian. In addition, the development and evaluation of objective procedures for analyzing internal evidence[1] is a challenging problem for the scientist or statistician. Several recent and important quantitative studies have focused on common function words, using text samples of known authorship to evaluate both potential authorship markers and statistical procedures. The goal of this study is to test the effectiveness of such an approach in the field of Jacobean drama. The problem chosen is the question of the possible collaboration between William Shakespeare and John Fletcher in *The Two Noble Kinsmen* and *Henry VIII*.

Perhaps very few literary scholars today would not accept the premise that the quantitative analysis of textual features is indeed a valid approach to solving authorship problems. Some argument may arise, however, concerning the

---

[1] Textual scholars use the term *internal* evidence to refer to stylistic or linguistic features of a text that may indicate authorship. *External* evidence refers to information outside the text itself that can be used to attribute the work (for example, details of printing, title-page ascriptions, references to the work in catalogues or registers).

relative weight of such assessments when compared to the results of traditional scholarship. Often quantitative statistical methods are developed by those who are dissatisfied with the subjectivity inherent in stylistic studies. While literary scholars often recognize the element of subjectivity in their work, they are understandably suspicious of procedures that make no acknowledgement of their own expertise and methodology.

I believe that traditional literary approaches and the scientific methodologies that are now being developed can and *should* be used to complement each other in a study of authorship. A scientific study of internal evidence is based completely on the data that a text offers; a researcher embarking on such a study cannot afford to ignore the information scholars have discovered about the text in question. (In a Jacobean problem, for example, a scientist who does not consider textual origins or external evidence can easily attach significance to textual features that are not the author's but instead are the results of scribal transmission or the printing process.) On the other hand, the statistician Kemp's observation (in "Personal Observations on the Use of Statistical Methods in Quantitative Linguistics" [59]) seems equally reasonable: "Judgements which take no account of quantitative evidence are as susceptible to criticism as methods of discrimination entirely based on such criteria." It seems obvious that any argument should be based on as much evidence as possible, and in an ideal world the results of both types of study would be used to support each other's findings.

The field of Elizabethan and Jacobean drama presents an investigator with major difficulties along with some most interesting questions. Any dramatic text is a challenge to stylistic studies. The text of a play reflects spoken language, sometimes representing natural conversation and often marked by a less rigid syntactic structure than literary prose. A good dramatist distinguishes his characters from one another through their use of language, and anyone wishing to study the playwright's own style must cut through his many different voices to discover common patterns. English Renaissance dramas may be more complex than most modern plays since they are written in verse (for the most part).

Dramatic creation in the Elizabethan and Jacobean theaters was very much a commercial rather than a literary activity. The texts themselves were generally not regarded as having value as works of literature; they were therefore subject to alteration or revision by the author or the company during production or revival. Schoenbaum, in *Internal Evidence and Elizabethan Dramatic Authorship*, goes as far to say that all plays "are in a sense collaborations, shaped from conception to performance by the author's awareness of the resources of actors and theater, the wishes of impresario or shareholders, and the tastes and capacities of the audience" [125, pp. 149–150]. Finally, the surviving versions of many plays may contain alterations introduced by scribes, prompters, editors or compositors.

Nevertheless, it is likely that features in the surviving texts reflect an author's subconscious habits of composition and that some of these can be used to distinguish him from other playwrights. The problem of identifying such traits is simplified when there are only two candidates for the authorship of a play or parts of a play. The possibility of collaboration between Shakespeare and Fletcher in *Henry VIII* and *The Two Noble Kinsmen* has been a much-debated issue and is the subject of the procedures described and developed in this study.

1.1 A Brief Description of the Authorship Question

Much of the controversy surrounding the possibility of collaboration between Shakespeare and Fletcher stems from the existence of *The Two Noble Kinsmen.* John Waterson's 1634 entry in the Stationer's Register attributes the play to both men, and the title-page of the quarto he published later that year also lists the two men as authors. While title-page ascriptions cannot always be trusted, support for the collaboration theory gained support in the 19th century. Many critics found contrasts in the two writers' styles and then identified these characteristics in different parts of the play. These judgements were often characterized

by the subjective excesses of *bardolatry;* Shakespeare was usually credited with any scenes considered to have merit, while the weaker and bawdier portions were given to Fletcher.

The continued discussion surrounding *TNK* made Victorian scholars more aware of the possibility of Shakespeare's collaboration with the man who succeeded him as the leading playwright for the King's Men. Perhaps it was inevitable that such a collaboration was suggested as an explanation of the stylistic and thematic ambiguities in a play for which no external evidence supporting collaboration exists, *Henry VIII.* An article by Spedding first explored this explanation in 1850 [152]; he starts from his critical dissatisfaction but supplements the argument with a table showing the proportion of verses that end in feminine (double) endings. Fletcher's fondness for feminine endings had long been noted, and Spedding shows that the distribution of these endings corresponds exactly to the scene by scene division which he proposes on stylistic grounds. Metrical tests such as this one proved to be the main weapon of the *disintegrators,* who zealously attempted to recognize texts (and parts of texts) in Shakespeare's accepted canon that were not really his. (The wild excesses of these scholars and the subsequent conservative reaction is thoroughly and scathingly documented in Schoenbaum's *Internal Evidence and Elizabeth Dramatic Authorship* [125].) Thus both *TNK* and *H8* were subjected to a battery of verse tests to support the division of each play between the two playwrights.

Some began to argue against Shakespeare's presence in either play, substituting another dramatist (usually Massinger) as Fletcher's partner. The controversy brought about by this contention helped to establish the details of Spedding's attribution as the orthodox collaboration view. Discussion now centered on whether the two parts of either play were indeed different and whether one of these could be identified as Shakespeare. These have been the goals of most of the authorship studies of these two dramas in this century, and scholars have made use of a wide-variety of textual and stylistic evidence. Parallel passages and metrical tests have been examined and Victorian metrical findings re-evaluated.

Studies based on the analysis of imagery and image-clusters have been presented, but such procedures have not been entirely accepted.

More impressive have been the results based on analyses of Shakespeare's and Fletcher's linguistic characteristics, including colloquial contractions (Farnham [35]) and philological innovation (Hart [46]). Very striking differences between the two writers' use of several pronoun and verb forms have been most convincing. Studies by Partridge [120] and Hoy [55] have been founded on Fletcher's preference for *ye* and *'em* instead of *you* and *them* and Shakespeare's use of the older inflectional ending for two auxiliary verbs, *hath* and *doth.* (These studies and others will be examined in more detail in Chapter 7.) Theories of additional major participants have been generally dismissed since the early part of the century. Questions remain regarding the nature of the relationship between the two dramatists in writing the plays: did they work in partnership, or did Fletcher take and alter a Shakespeare draft? In many ways this is the most interesting question and the least likely to be answered, certainly by objective methods.

While most textual critics have probably accepted that the texts of these two plays reflect the work of both Shakespeare and Fletcher, dissenters still exist. This is particularly true of *Henry VIII.* As Foakes notes in the introduction to his 1958 edition of the play [130], those supporting collaboration nearly always argue that the work lacks unity, while those supporting the theory of sole Shakespearean authorship usually admire the play. G. Wilson Knight expresses the point of view of the second group; after discussing "the play's artistic and organic validity" he returns to the authorship question:

> All this, I shall be told, would be well enough, if there were ten syllables in each line of the great speeches here and not eleven. Frankly, I do not know how satisfactorily to answer this objection: because I do not understand it. I believe such pseudo-scientific theorizing is again here, as elsewhere, merely an unconscious projection of our sense of organic incoherence within the play due to failure in focus and understanding [64].

From this quotation it is clear that the issue of authorship is not simply of historical interest but has fundamental consequences for the interpretation of *H8*. While proponents for sole authorship are on the defense, the issue is still considered uncertain enough that Bevington, in his 1980 edition of Shakespeare's works [126], falls back to the conservative position. "We are safest in assuming that the nineteenth-century efforts at disintegration are now happily out of fashion, and that the Folio editors knew what they were doing when they included *Henry VIII*." While recognizing Shakespearean elements in *TNK*, he omits it from his edition because "as a whole it seems considerably more Fletcherian than Shakespearean."

So the question of the authorship of individual scenes in *Henry VIII* and *The Two Noble Kinsmen* is an excellent test for evaluating the effectiveness of stylometric methods for solving Jacobean authorship problems. A large number of plays by the two candidates that are suitable controls have survived (perhaps more than in any other Elizabethan or Jacobean authorship question). While the exact nature of the collaboration remains cloudy, critics have accepted that (for the most part) individual scenes are by one man or the other. There is strong linguistic and stylistic evidence supporting existing attributions that stylometric results can be measured against. At the same time, the 19th century scene attributions have often been accepted as a whole without any re-evaluation by scholars attempting to demonstrate the presence of an author's hand. Finally, some scholars are unconvinced by the evidence for collaboration, which makes this a more interesting study for both this student and (one hopes) the reader.

1.2 Stylometry

Outside the areas of teaching and the preparation of textual apparatus, computing in Shakespeare studies has focused on analyses of style and authorship. One common term for this area of study is ***stylometry***. The recently published supplement to the *Oxford English Dictionary* defines stylometry as: "The technique

of making statistical analyses of the features of a literary style, esp. by means of a computer." The combination of the word *style* with the suffix *-metry* signifies the action, process or art of measuring style.[2] The first citation given in the *OED* supplement dates from 1945. In their 1973 paper "Positional Stylometry" [99], Michaelson and Morton note that the problems addressed by stylometry are usually questions of authorship, integrity or chronology: in other words, questions relating to the definition and description of an author's canon. Holmes' recent article entitled "The Analysis of Literary Style — A Review" provides a thorough review of the many approaches to stylometry developed in the last fifty years [51].

The central concepts of a statistical approach to authorship study have been recognized at least since Mendenhall's examination late last century of the claim that Bacon wrote the works attributed to Shakespeare [85]. By comparing the variation within samples of an author's work to the differences between writers, textual features are isolated that can be used to discriminate between writers. The recognition and analysis of sample variation distinguishes the stylometric approach from many quantitative studies of textual features by literary and linguistic scholars. Arguments for collaboration in *TNK* and *H8* have often been supported with evidence showing differences within each play, but the degree of internal variation found within Shakespeare's and Fletcher's unaided dramas is rarely presented for comparison.

It is not entirely obvious how best to analyze the significance of internal variation. To evaluate a proposed set of authorship markers, one researcher recently divided *The Winter's Tale* into two parts according to 1716 different combinations of scenes. He performed a statistical analysis on each division,

---

[2]The word *stylometrics* was often used in the late 18th century to describe verse-ending tests of authorship. (*Metrics* is the science or art that deals with meter and versification.) In this study "stylometric" will be used as an adjectival form for "stylometry." The *OED* supplement also lists the term *stylostatistics*, "the application of statistical methods to the analysis of features of literary style." The first recorded instance is from Herdan's 1956 book, *Language as Choice and Chance*, but in recent literature this word seems to be used less frequently than *stylometry*.

demonstrating that the difference between the parts of *Henry VIII* was no greater than that found in one of Shakespeare's unaided works.[3] A preferable approach to determining the two writers' shares in *TNK* and *H8* is the development of a procedure that allows each scene of the texts to be individually tested and attributed. Of course, there are a number of difficulties with such an ideal approach, but the ability to compare and assign small samples on an individual basis is one goal of the present study.

1.3 Choice of Variables

Tallentire outlines two possible ways of proceeding when using statistics to solve any literary problem [158]. In the first, an investigator subjects control texts to various tests to determine if statistically significant differences are present "with respect to *arbitrarily* chosen parameters." The second approach stresses that statistics can provide an objective component of judgement when a researcher makes hypotheses about characteristic textual features that he has noted through careful reading. The statistical study of the authorship of *Henry VIII* and *The Two Noble Kinsmen* described in this dissertation reflects the first approach.

Judging from comments heard at a discussion of stylometry at the 1986 conference of Association for Literary and Linguistic Computing, literary scholars seem to be more receptive to the second approach. "Count things that make sense," was one participant's plea. A statistical analysis of meter or sentence-length is based on textual features that are familiar to a person traditionally trained in the study of style. But the first approach is equally valid if the tests under examination have been validated in a sufficiently large number of control samples. When no apparent logical explanation can be proposed to explain why a test should indicate authorship or date of composition, a scientifically trained

---

[3]This study by Smith will be discussed in greater detail in Section 3.3.3, which starts on page 121.

mind can easily accept that the validation procedure justifies the method; a person with a non-scientific background will probably feel less comfortable about such procedures and their results.

Literary scholars may mistakenly accept quantitative studies of "sensible" textual features when the differences or similarities are not actually significant. The history of the study of parallel passages as authorship tests in Elizabethan plays is well-documented by Schoenbaum [125]. Parallel passages have an intuitive appeal to a literary mind, but in a number of cases "negative checks" have shown that supposedly significant parallels in two texts are not peculiar to the writer who has been proposed as the author of both. Unfortunately the textual features that stand out to a literary scholar usually reflect a writer's conscious stylistic decisions and are thus open to imitation, deliberate or otherwise. Tests of authorship that are founded on subconscious habits are a desirable goal in most (if not all) applications.

Stylometric studies of the first type described by Tallentire should be able to avoid the pitfalls which caused the failure of many studies based on parallel passages. Variables and tests are not really "arbitrarily" chosen but are selected solely because they discriminate effectively. A recent example of such an approach is Ledger's cluster analysis study of the frequency of letters that make up words in Greek texts [68]. Anticipating objections and defending his approach, Ledger notes:

> The assurance that statistical methods (strictly speaking, the methods of MVA [multivariate statistical analysis]) are capable of confirming what we already know, is the only basis from which stylometry should proceed into the unknown. Too often this point has been ignored in stylometric research.... Doubt may also arise in the minds of some who dislike the heavy mathematical content of these methods, and the difficulty of relating the final discriminant function, or group of functions, to any recognizable linguistic phenomena. In defense, it may be said that the mathematics helps to provide objectivity, and that a theoretical relationship most certainly does exist between the mathematical end product and the underlying patterns of linguistic

> behavior, though this relationship spans several levels of complexity and may be difficult to define in practice.

Ledger's last point emphasizes that the complexities of the relationship between the mind and language are not fully understood. The description of stylometric authorship tests as "verbal fingerprints" is often made. The validity of fingerprint identification was accepted long before the phenomenon could be explained. (The parallels between the current state of stylometry and early efforts of scientific identification systems, such as Bertillion's system of physical dimensions and fingerprinting, have been described by Morton in *Literary Detection* [108].)

A number of researchers have suggested characteristics that variables in stylometric analyses should possess. Borrowing ideas from a study of painting, Damerau [26] notes that good markers of personal style style will, first, not be prominent, in order to avoid imitation. Second, they should result from "mechanical execution;" it is taken that this will result in low variation from work to work. Third, they cannot result from convention and must show large variance when compared to the work of others. Finally, their frequency of occurrence should be high when compared to the sampling error. These views are echoed by Bailey, who asserts that useful variables should be "salient, structural, frequent and easily quantifiable, and relatively immune from conscious control."[4] Morton adds that an authorship study ideally requires habits of composition that are unaffected by variations of literary form, date of composition and subject matter [108].

Many studies have focused on what are often called *function words*. This very general term usually embraces prepositions, conjunctions, articles, pronouns, and some adjectives, adverbs and auxiliary verbs. These words satisfy the suggestions for stylometric variables in many ways. Because function words are required to construct most statements in a language, one can imagine that tests

---

[4] "Authorship Attribution in a Forensic Setting," in *Advances in Computer-aided Literary and Linguistic Research*. Edited by D. E. Ager, F. E. Knowles, and M. W. A. Smith; published 1979. Quoted from Holmes [51].

based on these forms may often be insensitive to changes of subject and style. They are frequent; the most commonly-occurring words belong to this category. Although the most common function words are necessary for composition, a writer is in many ways not conscious of how they are being used. Their rate and pattern of occurrence may therefore be relatively immune to conscious stylistic manipulation or imitation. In most cases, function words are easily counted (by computer, at least). Function words have featured prominently in a number of successful authorship studies (most notably, in Mosteller and Wallace's analysis of the disputed *Federalist* papers [113]).

1.4 Statistical Inference

To understand how statistical methods are used in an analysis of authorship, one must explore some of the foundations of statistical theory. The field of statistics can be broadly divided into three roles: description, inference and prediction. Prediction does not usually play a part in literary statistics although it is important in commercial situations. The role of descriptive statistics is to summarize and condense large amounts of raw data while preserving as much significant information as possible. The statistical term "inference" must be distinguished from its colloquial counterpart. In statistics the term is always associated with the idea of *samples* and *populations*. Snedecor and Cochran, in their introduction to statistics [151], define statistical inference as the inductive "process of making statements about the population from the results of samples." Inferential techniques allow a statistician to make well-founded statements about the population when it is impossible or impractical to observe the entire population. Indeed these procedures make it unnecessary to study the complete population; given information from the samples, inference allows one to make statements about the characteristics of the entire population, with a given probability of accuracy.

In the context of Shakespeare and Fletcher's collaboration, the population can be defined as all of the drama written by Shakespeare. This population can never be exactly determined since certain texts may have been lost or unfinished, and particular questions of collaboration may be beyond resolution. Likewise one can postulate a population of Fletcher's dramatic writings. Using *statistics* calculated from samples known to be from these respective populations, one can determine the value of *parameters* for a population within given confidence intervals. ("Statistics" is used here as a plural noun, distinct from its use as a description of an area of mathematics.) Data from an unknown sample (such as a scene from *Henry VIII*) can then be compared to the parameters of the two populations to see how closely it resembles either one. The conclusions reached at each of the various stages of this process are based on statements of probability.

Samples, populations, statistics, parameters and statistical inference are basic fundamentals of statistics and will be covered in any introductory textbook. Humanities scholars who are unfamiliar these ideas may wish to read Kenny's introduction to literary statistics, *The Computation of Style* [61]. A more advanced discussion of statistical subtleties can be found in Thomson's series of articles in the *ALLC Bulletin* subtitled "On the Small Print of Statistics" [161]. Thomson and Kemp [59] discuss the problems of definition concerning one basic statistical term of great importance in authorship studies: probability.

1.4.1 Interpretations of Probability

Even those with little statistical background understand the interpretation of probability behind the statement: "The probability of rolling a three with a fair die is one-sixth." Those with some statistical learning will probably have little problem interpreting a statement such as: "The probability is 95% that measurement X of sample S will fall between 7.3 and 12.5." But what of the statement: "The probability that Shakespeare wrote all of *Henry VIII* is less than 100 to 1"? Clearly Shakespeare either did or did not. The *relative frequency*

definition of probability that we learn in school is difficult to interpret in this context. Does one view this as a single event (like one toss of a coin)? If so the probability should really be either one or zero. Or do we begin our interpretation of the statement with a hypothetical proposition: "Given 100 plays with the same characteristics as *Henry VIII...*"? Clearly an interpretation of *degree of belief* is involved here, but how does statistical theory accommodate what appear to be two different interpretations of such a fundamental concept as probability?

A browse through the introductory chapters of books on probability theory reveals a great debate regarding the definition of probability, a concept which underlies all statistical theory. Thomson describes this as "the meta-scientific question of relative frequency v. subjective belief," but in *The Foundations of Statistics*, Savage outlines three main views, which he summarizes as follows [124, p. 3]:

Objectivistic: (sometimes called "frequentist" or "statistical") Some repetitive event (such as tosses of a penny) proves to be in reasonably close agreement with the mathematical concept of independently repeated random events. This is the long-run relative frequency view.

Subjective: (sometimes called "personalistic") Probability measures the confidence that a particular individual has in the truth of a proposition. The view assumes that the individual in question is "reasonable;" it also recognizes that two such individuals may have different degrees of confidence given exactly the same evidence.

Necessary: A view that regards probability as an extension of logic. Probability measures the extent to which one set of propositions, out of logical necessity and apart from human opinion, confirms the truth of another.

The necessary view is reflected in works by Keynes [62] and Jeffreys [57] published before 1950. Savage writes that in 1954 the majority of statistical researchers in the English-speaking world adhere to some form of the objectivistic viewpoint.

This group, the British-American school (as Savage calls it), is responsible for most of the advances in statistics this century. While the success of procedures developed by advocates of this view might be taken as evidence that the underlying assumptions regarding probability are correct, many objectivists recognize a problem. According to their theories such statements as "there's a 30% chance of rain tomorrow" may have a meaning, but this meaning is not relevant to the mathematical concepts of probability [124, p. 62]. Indeed von Mises (who might be viewed as a "fundamentalist" among objectivists) clearly states on the first page of his *Mathematical Theory of Probability and Statistics* [165] that probability theory has nothing to do with such questions as the likelihood that two countries will soon go to war, or that the *Odyssey* and the *Iliad* were written by the same author. These deal with particular situations, and the probability theory from which he derives his comprehensive theory of statistics is limited to "a mathematical theory of repetitive events."

This limitation seems rather severe. It is desirable to include such statements about the chance of rain or the Homeric problem within the framework of statistical theory. More crippling is the implication that one cannot choose the most promising course of action from a number of possibilities by statistical methods; the theory applies to events and processes but not to propositions [124, p. 4].[5] Even the assumptions that lead to the relative frequency model may rely on subjective beliefs. Thomson [161, Part 1] describes a line of argument showing that the idea of an infinite series of independent random events conflicts with the idea of a mathematical limit; thus, any probability based on this model involves a statement of reasonableness regarding predicted results. The adherents of the subjective school maintain that they can derive a theory of statistics which is consistent with the many successes of the objective approach. Yet such a theory centers on subjective belief by an individual.

---

[5]The concept of decision is central to Savage's own subjective view; he argues that any satisfactory account of probability must deal with the concept of action in the face of uncertainty [124, p. 60].

De Finetti states his initial thesis in large, bold capitals: "Probability does not exist." The two volumes of *Theory of Probability: A Critical Introductory Treatment* that follow this startling assertion attempt to demonstrate that statistical theory survives without it. In a spirited introduction, he maintains that probability, if regarded as having some kind of objective existence, is no less misleading than notions of the cosmic ether, absolute space and time, and fairies and witches. The objectivistic view of the concept would be "an illusory attempt to exteriorize or materialize our true probabilist beliefs" [27, p. x].

If the subjective view does indeed replace the relative frequency idea, then one's fundamental conception of probability must be altered. Science proceeds by developing models that describe natural phenomena, and as scientists learn more these models are refined or rejected. Some models have been so successful that some people have begun to believe that the model really is a complete and accurate description of "natural law." In such cases major new advances that show an established model's inadequacies appear disturbing to many at first (but then often capture the imagination). The theory of relativity has thus affected our perceptions of time and space; so, to a lesser extent, has quantum mechanics affected our views of matter. Perhaps the subjective view of probability represents another such upheaval, but for the present the issue appears to be unresolved.

What are the implications of this controversy for those who wish to use statistics in attribution problems? The debates of theoretical statisticians do not undermine our confidence in any standard statistical procedures. Much of de Finetti's work is aimed at building a bridge connecting the new subjective approach to the results that stem from the objectivist ideas. Whatever probability actually means, everyone agrees on when probabilities are combined by multiplication and when they are added together. Savage notes: "Considering the confusion about the foundations of statistics, it is surprising, and certainly gratifying, to find that almost everyone is agreed on what the purely mathematical properties of probability are" [124, p. 2]. Thus the principles and procedures

developed in the last fifty years are accepted by both schools of thought. But it should be noted that the distinction drawn by Kemp [59] between "probabilities" and "measures of degree of belief or conviction" may not not be a valid separation if the views of Savage and de Finetti are correct.

1.4.2 Measuring Differences

By examining the debate about probability one is forced to question the basic reason for using statistical techniques in a literary problem. Statistics' role is to provide an objective *measure of the difference* between samples of text. Past research has usually expressed this measure in terms of probabilities. One may justifiably ask "Why probabilities?" given the questions of interpretation surrounding this term, especially in regard to propositions. Probabilities have proved to be extremely useful in analyzing sampled data in many other fields: gambling, biology, agriculture, insurance, to name but a few. But probability is intimately associated with the idea of random occurrences, which seems very much at odds with our intuitive feelings about language. One does not think about composition as a random process. Whatever measure of difference we choose, straightforward and acceptable procedures should exist for combining results for different variables in one comparison. Probability certainly has mathematical properties (familiar to most readers) which specify when and how one can combine independent probabilities. But Chapter 3 will describe how difficulties in combining significance test probabilities in some authorship studies have led to controversy.

One can think of other natural ways of expressing differences. Distance is such a measure, usually expressed by scientists in terms of meters: not a very intuitive unit for the study of literary features. But in comparing plays "Shakespeare and Marlowe are miles apart" sounds just as reasonable as "There's a 1000 to 1 chance that Marlowe wrote Shakespeare's plays." Distance measures are a more universal concept than one might have imagined 50 years ago. They are particularly attractive in multivariate situations, since distances between points

in multidimensional spaces are easily calculated. By representing samples as *vectors* defined by the values of many different variables, one can obtain an objective measure of dissimilarity by calculating a multidimensional distance measure. Questions of interpretation can still arise, particularly in regard to units and scaling when variables are different in nature. The last two decades have seen great developments in *cluster analysis* and *discriminant analysis*, statistical procedures that make use of distance metrics for the multivariate comparison of samples. Although discriminant analysis procedures produce a probability as a final result, the combination of results is not a problem because the method is truly multivariate. Chapter 6 will discuss these ideas further and evaluate the use of distribution-free discriminant analysis in this authorship problem.

1.4.3 Justifying Interpretations and Procedures in a Literary Context

The overall caution expressed by Kemp in the discussion about probability [59] seems well justified. Commenting on the debate he argues "that it is better to defer such questions of definition and related philosophic issues until there is evidence, from specific studies, that statistical method fulfills a useful function in the analysis of literary characteristics amenable to measurement." Inferential statistical procedures must be applied in a number of controlled experiments in order to assess the appropriateness of a procedure for the analysis of textual features. In another article, published in the *ALLC Bulletin* [58], Kemp comments that it does not automatically follow that procedures appropriate to agricultural experiments are equally valid for establishing the chronology of some texts. Likewise "without due consideration" one cannot justify the use of decision rules designed for determining which of two drugs is more effective in resolving authorship problems.

Regarding the use of particular statistical procedures in authorship studies and the interpretation of resulting probabilities, some argue that "the proof of the pudding is in the eating." The study of the statistical characteristics of literary

features is still in its infancy, however; a great deal of eating is still required before anything can be accepted as "proved." Few if any procedures for resolving authorship questions have been accepted as being of general usefulness since Mendenhall's initial analysis of word length in English, published one hundred years ago.

In this study of the collaboration of Shakespeare and Fletcher, less emphasis will be placed on probabilities of authorship than on the establishment of methods that reliably classify known samples. In this context *classification* is the process by which a sample is determined to have been written by one or the other of the two writers. While obviously one would like to discover strong evidence one way or the other, discrimination procedures should not be evaluated according to the magnitude of the probabilities established for a small number of control samples. A test that attributes *The Tempest* to Shakespeare with a likelihood ratio of one hundred million to one may fail to classify 25% of the other plays of undisputed authorship. An approach based on misclassification rates requires that a large number of machine-readable texts be available and that facilities exist for the efficient processing of large amounts of data. If this study succeeds in establishing more about the effectiveness of stylometric attribution in English Renaissance drama, it will be due to the fact that both these conditions could be satisfied.

1.5 Overview

As stated at the beginning of this introduction, the subject of this dissertation is the use of computers and statistical methods in the study of questions of disputed authorship in the field of Shakespearean drama. To demonstrate the effectiveness and limitations of stylometric techniques, I have chosen to examine the question of the possible collaboration of Shakespeare and Fletcher in the composition of *Henry VIII* and *The Two Noble Kinsmen*. The study will concentrate on tests based on the occurrence of function words and frequent word classes. A

secondary objective of the study is to make the most of computing techniques to process texts efficiently and to facilitate the statistical analysis.

The first step in a stylometric study is the choice of texts used to establish the characteristics of each writer's composition. When considering the plays of Shakespeare and Fletcher, such decisions are by no means simple. A number of difficulties arise even before considering the availability of machine-readable texts. One must consider whether modernized or old-spelling editions are preferable for use in an authorship study (while recognizing that the existence or non-existence of either may leave one no choice). For Shakespeare I have chosen to use machine-readable editions of the original quarto and folio texts. For Fletcher, existing versions of several early authoritative texts have been supplemented by four critical old-spelling editions. Chapter 2, entitled "Text Selection and Processing," first outlines the reasons behind these decisions and describes some characteristics of old-spelling texts that influence a stylometric analysis. This chapter then describes the set of control texts selected and justifies the more difficult choices. (Naturally the sources of the early printed editions and the transcription and printing process must be considered in choosing control texts and methods of analysis. Appendix A provides a brief description of the sources and printing of 17th century texts for those unfamiliar with the field of Shakespearean textual studies.)

Chapter 2 also provides details of the computer processing of these texts. The format of the texts stored in the computer files is described. Some modifications and light editing of the texts was necessary: abbreviations were expanded, spelling conventions regarding *i-j* and *u-v* were made consistent, and some word divisions were adjusted. A system of marking words in the files was developed to resolve problems of variant spellings and homonyms, and a program was used to standardize the variants of high-frequency words. A similar system of coding and computer processing was used to expand compound contractions (such as *it's* and *'tis*) to their full forms with a minimum of manual labor. These sections of Chapter 2 also describe some of the computing methods used to search for

features in these texts and store the resulting counts. The chapter closes with a quantitative analysis of compound contractions in each dramatist's texts.

Chapter 3, "Some Recent Stylometric Studies," reviews a number of stylometric authorship studies. Morton, Michaelson and their associates have developed and evaluated a number of attractive general tests of authorship for English texts. One of the demonstrations in "To Couple Is the Custom" [102] was an analysis of the homogeneity of *Pericles*; this study sparked off considerable interest and controversy in the field of Jacobean stylometry. Chapter 3 describes the development of these methods and reviews how they have been applied to various authorship questions by Morton and others. Other researchers' criticisms and refinements of both the variables and the procedures are also reviewed in this chapter. Other stylometric studies of Elizabeth and Jacobean questions are also reviewed, along with the few applications of discriminant analysis in authorship studies.

Chapter 4, "Collocations and Proportional Pairs," describes an analysis based on two of the tests of authorship developed by Morton and his associates. Collocations and proportional pairs have been used in a number of studies by Morton and others, and are the most appropriate of the tests described in "To Couple Is the Custom" for Shakespearean problems. Thirty frequent collocations and five proportional pairs are examined in plays of known authorship. Results presented in this chapter demonstrate the limited usefulness of χ^2 tests for examining within-author variation. The variability of these 35 features within each author's control set is large enough that methods based on these features are not [illegible] effective in discriminating between samples of the two writers' works.

A number of individual words and some word classes are shown to occur in the two writers' texts at significantly different rates. Most of these occur more frequently than the collocations and proportional pairs examined earlier. Chapter 5, "Finding Common Words that Discriminate," describes how the rates of occurrence of individual words and groups of words were examined in the

control texts to discover which are the best authorship markers. Changes in these rates due to genre and date of composition are investigated.

Chapter 6, "Discriminant Analysis of Word Rates," introduces the multivariate statistical technique of discriminant analysis. Two distribution-free methods, kernel estimation and nearest neighbor classification, are described in detail. Feature selection methods are used with the kernel method to find the most effective subsets of the marker words isolated in Chapter 5. The effectiveness of the classifiers is evaluated for text samples of varying length, for samples made up of the speeches of individual characters, and finally for samples composed of text by both writers. A *reject option* is used to recognize scenes that cannot be safely assigned by the procedure, and rejected and misclassified scenes are examined in detail.

Chapter 7, "Applying the Classifiers to the Disputed Plays," describes an analysis of *Henry VIII* and *The Two Noble Kinsmen*. The textual history and external evidence for authorship for each play are briefly examined, and past studies of authorship based on the quantitative analysis of linguistic features are reviewed. The discriminant analysis procedures developed in the previous chapter are applied to scenes from the two plays. The rate of each marker in a scene is examined independently in an effort to determine how individual words might affect the multivariate classification result. The dissertation concludes with a discussion of the study's procedures and results, and indicates some areas for further research.

Chapter 2

Text Selection and Processing

To begin an authorship study one must first select a set of texts and editions. This chapter will describe the texts of Shakespeare and Fletcher chosen for use in this study. Versions of diplomatic and old-spelling critical editions have been used. Problems associated with this choice are discussed and the choice is justified on the grounds of reliability and availability. For each author a number of plays have been selected from those that are free from serious textual problems to act as a *control set* for comparisons between the two playwrights. Other texts with some textual difficulties have also been used due to availability in machine-readable form or because there were good reasons for wanting to make use of them. These texts (and some others without any problems) make up a *test set* used to validate the authorship methods tested in this study.

The second part of this chapter discusses details of computer processing of these texts. Details of the data files are described, and a coding system based on *hash suffixes* is introduced for distinguishing homonyms and recognizing variant spellings. A number of common function words frequently occur in contracted forms, and the coding and replacement strategy has been broadened to allow all contracted forms of common words to be expanded.

2.1 Modernized or Early Editions?

In some studies of authorship problems the choice of texts will be relatively straightforward; only a few authentic texts by the candidates may exist and a particular edition may be generally accepted as a standard. But a large number of options confront the person beginning a study of Jacobean plays. A number of modern editions of Shakespeare's works are available, but many other plays have only been published in old-spelling editions. Such texts present a number of problems for the investigator, especially if a computer is to be used to find and count features in them. This would suggest that modernized editions should be chosen, but many scholars argue convincingly that authorship studies should be based on the earliest authoritative texts that have survived.

Past researchers have spent little time justifying their choice of texts, yet quite often the linguistic variables they measure depend a great deal on features that may have been altered by an editor. At one extreme there are editions that merely try to reproduce the early document. Such editions are known as *diplomatic texts;* perhaps the best known examples in the field of English Renaissance drama are the Malone Society Reprints. When an editor modifies readings from the original authoritative manuscript or printed edition he produces a *critical text.*

Principles of editorial practice have developed a great deal in the last fifty years, and each modern critical edition can reflect various theories and approaches. Bowers, in *On Editing Shakespeare and the Elizabethan Dramatists* [15], views the task of editing a Renaissance dramatic text as an attempt to synthesize the one or more authoritative early documents with "hypothecated" readings from the author's manuscript, which is usually lost (p. 75). The critical text that results is thus eclectic whether or not it attempts to modernize spellings or punctuation. The debate about the relative merits of modernization and old-spelling has not been resolved. Clearly students and actors benefit from

a modernized edition, and the descriptions of discussions at the 1978 Gleneldon conference on old-spelling editions reveal that even the advocates of old-spelling sometimes question their own position [139]. Despite one claim at the conference that old-spelling editions are only for old-spelling editors, many of the participants still agreed that such texts better preserve authorial intentions and provide the textual scholar with more accurate information for analyzing a play.

While modern editions of Shakespeare's works are published every decade or so, other Elizabethan and Jacobean dramatists suffer from neglect in this regard. The most recent modern editions of some canons date from the early decades of this century and do not reflect the considerable advancements of knowledge regarding 17th century texts since that time. (This is the case for the works of Beaumont and Fletcher; two different editions were published between 1904 and 1912.) While some plays have been treated in old-spelling critical editions, others have only appeared in diplomatic form. This is the case for many of the obscure or corrupt texts that are often at the heart of authorship controversies.

Despite the large number of modernized Shakespeare editions, no old-spelling critical edition of his complete works has been available, a fact which Bowers [15, Note 35, pp. 124–125] feels has seriously hobbled Shakespearean criticism. The long absence of such an edition can easily be explained by consideration of the textual problems posed by the large canon, commercial considerations, and the high academic stakes involved. This situation has changed recently. In November 1986 the Oxford University Press published a critical old-spelling edition to accompany the new one-volume modern edition of Shakespeare's works. Unfortunately this edition appeared at the end of my research and was not used in this study.

2.1.1 Availability of Machine-Readable Texts

The extent to which a computer can be used in a study obviously depends on the availability of machine-readable texts. The Oxford University Computing Service's Text Archive has one of the largest collection of machine-readable

texts which includes approximately one hundred Renaissance dramatic texts. Of these the only modernized editions it included in the spring of 1983 were *Julius Caesar* and the works of Christopher Marlowe. Most of the others are apparently diplomatic versions. A machine-readable version of the modern Riverside edition of Shakespeare's works [136] (edited by Evans) does exist but is not generally available to researchers.[1] In the first edition of the journal *Literary and Linguistic Computing* (published Summer 1986) Ule has advertised the availability of fifty transcriptions of original Elizabethan texts with standardized American spelling. While the possibility of complete standardization of spellings is an open question (this is discussed below on page 29), it seems likely that these files could be used in an authorship study based on common words.

The lack of modern editions in machine-readable form could certainly be rectified by having them all prepared or read (and certainly will be as new editors of the plays make use of word-processing and electronic publishing). But considerations of availability alone might justify the development of methods that do not require modernized versions. Clearly it would be desirable to develop techniques that could make use of the many existing old-spelling texts that already exist.

2.1.2 Reliability

Other reasons for working with early printed editions of texts, other than the lack of convenient modernized editions, have been brought forward. Schoenbaum gives several examples of past researchers who have placed significance on features in a text that are not found in the original version upon which all subsequent editions have been based [125]. One of the eight principles he outlines

---

[1] A machine-readable version of this edition was prepared by Spevack and used to create *A Complete and Systematic Concordance to the Works of Shakespeare* [155]. Responding to an enquiry about the availability of the files, Spevack wrote to me on 16 September 1981: "Since so many people here are engaged in Shakespearean research, it has been my policy not to duplicate the tapes."

for "avoiding disaster" is: "The investigator must always work with the most reliable texts, preferably directly with the early prints or manuscripts." Greg also strongly supports this view:

> It is time to recognize that an edited text — perhaps legitimate as an aid to aesthetic enjoyment — is from the point of view of every sort of critical investigation merely a text from which most of the relevant evidence has been carefully removed. To rely on it is like trying to solve an archaeological problem, not by the study of the finds *in situ,* but from neatly ticketed specimens in a museum.[2]

One must consider that Greg's remarks were made at the end of period when editors emended (often silently) a great number of features in a text, often only according to the editor's own preference. Since then editors have become much more conscientious in preserving the substantive (and to some extent the accidental) readings of their copy-texts and in informing the reader of any departures.[3] Bowers notes that critical, old-spelling editions are becoming the norm "for ordinary close literary study by an informed person" [15, p. 69]. Certainly he is somewhat responsible for the trend, since this modern approach is perhaps best exemplified by the recent editions of the works of Dekker and of Beaumont and Fletcher prepared under his general direction.

In the introduction to the multivolume series of the works conventionally attributed to Beaumont and Fletcher, Bowers states that one of his goals was to offer critical old-spelling texts "addressed principally to those who need to make a close study of the most minute formal characteristics of a text" in order to allow a textual critic to reconstruct the readings of the original copy-text in all essential detail [8]. In this series the editors have retained the accidentals of the early printed editions, but have corrected obvious typographical errors, expanded abbreviations and made speech prefixes consistent throughout a play.

---

[2]Quoted by Schoenbaum [125, p. 172], from *Modern Language Review,* XX (1925), p. 199.

[3]*Accidentals* include the general texture of capitalization, punctuation and spelling. *Substantive* revisions are verbal emendations of a more serious nature.

Each play's extensive critical apparatus carefully preserves details about the early copy-text. The extensive critical apparatus that accompanies each play helps make such texts almost ideally suited for use in authorship studies.[4]

Again, it is unfortunate for my purposes that such an edition of Shakespeare was not available, in printed or machine-readable form, at the beginning of my research. Diplomatic versions of the texts (both quarto and Folio texts) were prepared by the Oxford University Press Shakespeare Department for the series of old-spelling *Oxford Shakespeare Concordances* (described by Howard-Hill in an article *Studies in Bibliography* [52]) and eventually used to prepare the critical old-spelling edition. In order to attempt to evaluate authorship methods on a wide range of Shakespearean texts, I decided to make use of these available versions. Although these texts may not be as suitable for detailed study as critical editions, one can compensate for the dangers to some extent by understanding the problems associated with diplomatic versions of early texts and by applying this understanding to both the selection of control texts and the actual analytic methods. The original orthography certainly presents difficulties for computer processing. Coding techniques and programs can be used to handle many of the problems for high frequency words. Some of these problems will be examined in the next section, and the methods developed to solve them will be discussed in Section 2.5.3 on page 60.

2.2 Textual Considerations

There are several reasons why the textual problems associated with 16th and 17th century texts must be carefully studied in any authorship study of Shakespearean drama. Obviously in a study such as this, the disputed works must be compared against works of known authorship, but how certain can one be

---

[4]Indeed, Schoenbaum chastises two researchers who do not make use of the "superb Bowers *Dekker*" in their studies [125, p. 172].

that every word of a text such as *Macbeth* was written by Shakespeare? The earliest printed versions of the plays do not always accurately reflect the author's actual words. If an authorship method studies minute details of word choice or arrangement, then one must consider the possibility that the transcription(s) from the author's manuscript or the printing process itself may be responsible for the observed features. Also, if the researcher chooses not to make use of critical editions, then textual considerations must dictate the choice of text where more than one early version exists. Finally, many of the challenging authorship problems from this era involve plays that have only survived in seriously corrupt or defective versions. Thus, textual considerations must not be forgotten at each level of an authorship study: in choosing the texts to be studied, in the development and application of the actual tests, and in the evaluation of the results.

Characteristics of the texts themselves affect the *countability* of certain features. One must have a working definition of the occurrence or non-occurrence of a given word before one can measure features based on words. Clearly this would be a relatively minor problem in a modern prose text like this dissertation. However, the plays under consideration are dramatic, poetic and old; spelling variants and contracted forms are very common. In deciding how to treat variants and contractions one must decide whether occurrences of these forms can be assumed to be consistent and subconscious habits of writing and how safely one can attribute them to the playwright himself. The sources of early printed editions and manner in which the printing process may have altered these sources is described in Appendix A, page 345. Many of the decisions regarding variants, contractions and text selection have been based on findings outlined there.

2.2.1 Spelling variants

One of the most obvious problems of using a computer to study 17th century dramas is spelling variation found in these texts. If an authorship method relies

on the machine to find and count occurrences of certain words, then the computer, unlike a human, will see no relationship between the simplest of spelling variants and will treat the two forms as different words (unless special software is devised to recognize variants as such). Considerable variation of spelling between writers and within a single writer's works was a feature of Early Modern English even through Milton's time (see Barber [7], pages 15–23 and 114–121). The fragment of *Sir Thomas More* thought to be in Shakespeare's autograph provides a telling example: the word *shrieve* (a form of *sheriff*) is spelled five different ways in five different lines, and the name *More* is spelled three ways in a single line.

The compositors' work in the printing house has almost certainly complicated any attempts to determine an author's intended spelling. McKerrow, in *An Introduction to Bibliography for Literary Students* [82], provides evidence from the middle of the 16th century showing how compositors varied spellings in order to justify their lines of print (p. 11f). After studying several authors' manuscripts and their printed versions, he goes on to postulate a general rule ("for what it is worth"): a compositor would follow his own spellings in common words (and what he misread as common words) but would follow the spelling of the manuscript (or what he believed to be its spelling) in rare or nonce words (p. 249). Indeed, bibliographers rely in part on different compositors' spelling habits for common words to identify the output of a particular workman. (Much of our knowledge of the first complete edition of Shakespeare's plays stems from such an analysis by Hinman, *The Printing and Proof-Reading of the First Folio of Shakespeare* [49].)

But are all problems associated with spelling eliminated by the use of modernized editions? Certainly they are eased in regard to most of the common function words; all modernized editions would certainly normalize such different forms *been, beene, bene* and *bin,* or *do, doo, doe* and *dooe.* However, Ule's belief that the normalization "practice is established for modern spelling editions

of *Shakespeare*" [162, p. ix] is certainly untrue, as the examples in Wells' *Modernizing Shakespeare's Spelling* demonstrate [169]. As general editor of the new Oxford Shakespeare, Wells sets out to define a set of principles for modernizing spelling (a task treated as a mere "secretarial" problem by some editors) and to discuss the resulting editorial difficulties.

To begin with he notes that editorial practice regarding spelling differs considerably in different modernized editions. The Arden series varies from play to play. In his introduction to the Riverside Shakespeare, Evan's describes how he has preserved a selection of Elizabethan spelling forms from the copy-text in order to suggest a "kind of linguistic climate" by reflecting the original pronunciation of the time [136]. Examples of the variants he preserves include: *vild–vile, bile–boil, conster–construe* and *chevalry–chivalry.* Wells comments that this approach makes most editor's treatment of spelling variants seem "reckless" (p. 4).

Wells' own position is that a modern editor should not try to reconstruct the original pronunciation of a text and that nothing is to be gained from a hodge-podge of ancient and modern. He notes that no one has ever considered altering words whose spelling has remained the same but whose pronunciation has altered (as demonstrated by observed rhyming pairs; for example: *swan* with *can*; or *sate* with *bat, gnat* or *hat*). Wells also finds differences in how editors handle aphetic (for example *stonish* for *astonish*) and syncopated (for example *ignomy* for *ignominy*) forms, which are often varied to affect the poetry's meter.

Trickier problems arise with semantically significant variants; for example, the words *curtsy* and *courtesy* were represented in the 16th century by the same set of spelling forms, and only in modern times have we adopted the different spellings for the different meanings. Many editors have relied on the wisdom of the *Oxford English Dictionary* to determine if a spelling represents a distinct form, but Wells shows that the *OED* does not always make consistent distinctions

between spelling and form.[5] Clearly there is a wide variation of practice among modern editors which in part results from different editorial purposes. How this variation affects a method for determining authorship must be considered and measured when possible.[6]

2.2.2 Contracted Forms

The last quarter of the 16th century saw the beginning of a period when the use of contracted and weakened forms of words in drama became increasingly popular. Partridge, in *Orthography in Shakespeare and Elizabethan Drama* [120], examines these forms in a number of texts thought to preserve authorial characteristics. As mentioned in Chapter 1, contracted and weakened forms (such as *ye* and *'em*) have played a central role in past analyses of the authorship problem of *Henry VIII*. Indeed Partridge's own analysis of the play led him to carry out a comprehensive analysis of graphical forms in Elizabethan drama. His results are extremely valuable in determining how the common words at the heart of stylometry occur in contractions and in determining whether these forms might have been altered in transcription or printing.

Contractions involving just a single word are known as *simple contractions.* Two types, aphesis and syncope, were mentioned in the previous section in relation to differing modern editorial practices. Weakened forms such as *'em* for *them* are also familiar. *Compound contractions,* where two or more words are

[5] Wells expresses his views on this matter and gives more examples on pages 5–8 of *Modernizing Shakespeare's Spelling.*

[6] It seems to me that modernization of spelling is thoroughly ignored by many who advocate tests measuring vocabulary richness (such as those proposed by Muller and Ule). Ule's remarks in the introduction of his Marlowe concordance [162] are not very satisfactory, and his reliance on the practices of the editor of the Riverside Shakespeare seem to be very dubious in light of Well's recent publications [169,170]. Unfortunately, the existence of Hinman's concordance [155], based on the Riverside edition, means that those interested in Shakespeare's vocabulary will continue to rely on data from what many consider to be a poor edition for critical study.

coalesced in speech, occur in a large number of forms and often at a high frequency. There are several types of compound contractions. Two contractions of *it is* are common: *it's,* where the verb is *enclitic* (that is, the second word of the compound is pronounced as part of the preceding word); and *'tis,* where the *it* is *proclitic* (it has no independent pronunciation but is attached to the second word). More sophisticated forms include *initial assimilation,* where the proclitic word is reduced to the same letter that begins the second word and eventually is assimilated. One common form of this involves *he has,* as illustrated from *Demetrius and Enanthe:*

> *Line 2433:* I am vndon: has twenty Deuills in him...

End assimilation results from the same phenomenon with the enclitic word seeming to disappear: *this* < *this's* < *this is.*

The term *elision* (described by Partridge as "a word of comprehensive use and misleading connotation" on page 91) was originally used to describe the deletion of part of a word in order to reduce the number of syllables in a line of poetry. The term was later adopted to indicate a partial repression or slurring of a syllable. Obviously in some examples a single or compound contraction can satisfy both definitions. Although Smith [148] refers to occurrences of forms like *'tis* as "elisions," I will use "contractions" as a general term to refer to all these forms.

Why worry about contractions? Previous stylometric studies have treated contractions in different ways. Smith (in "An Investigation of Morton's Method to Distinguish Elizabethan Playwrights" [148]) notes that very different counts can result if one counts contractions as occurrences of the collocation represented by the full forms. His tables show that in Webster's *The White Devil* the 311 occurrences of *is* are followed by *the* 35 times (11.3%). If contracted forms are expanded the counts are 360 and 67 (18.6%). Middleton's *Women Beware Women* contains 148 occurrences of *is* in non-contracted forms and an additional 367 bound up in contractions. Smith remarks on page 7: "As there appears to be no general justification for assuming that an elision is other than a

bona fide occurrence of a feature ... their omission from Morton's and Merriam's counts would appear unjustified." Section 4.1.2 will demonstrate that the decision to expand or ignore compound contractions can have major consequences in collocation studies.

The choice to expand or not to expand involves more than a difference in working definitions. Partridge discusses at length (and seems to accept) [120, pp. 153–155] the conclusion of Farnham's 1916 paper [35] that the use of contractions such as *'tis, in't, i'th* and *let's* reflects the author's rather than the printer's habits in the texts of Beaumont, Fletcher, Massinger and Shakespeare. This conclusion does not appear to have been closely re-examined in light of the tremendous increase in our understanding of copy-texts and printing practices gained over the last two decades. Partridge himself gives a number of examples in earlier chapters which indicate that the author's intentions were not always preserved. Since a dramatist writing verse generally uses orthography to communicate a line's scansion to the actor, an examination of the meter may reveal changes introduced by scribes or printers. Partridge provides several examples from Shakespeare in which the use of contracted or full forms conflicts with the requirements of the meter.

He notes that full forms sometimes appear where the meter requires slurring in two early plays, *Romeo and Juliet* and *Richard II,* which he maintains are close to Shakespeare's autograph:

Rom 1299: She would be as swift in motion as a ball...

R2 675: Where words are scarce they are seldome spent in vaine...

Two weakenings ("probably not by Shakespeare himself" [120, p. 78]) are marked in *Richard II* where the meter requires full forms:

R2 472: Sweare by the duty that y'owe to God...

R2 531: Oh had't beene a stranger, not my child...

Remarking on some differences between the Folio and quarto versions of *Hamlet* (such as *i'th'* for *in the* and *I'm* for *I am*), Partridge concludes (page 101) that scribal modification reflecting an actor's stage performance "may, therefore, be

expected in First Folio texts based on late prompt copy." He also accepts this as a possible explanation for some Folio occurrences of *'em* and *ye:* not a reassuring conclusion in light of the importance of these two words in previous studies of Shakespeare and Fletcher.

Other examples in Shakespeare's plays can be found. Ure [132, p. xxviii] feels that there is "some reason to believe" that the compositor expanded such colloquial contractions as *I'll, he's, that's, o'er,* and *e'en* in the first quarto of *Richard II.* Taylor, in studying variants between the Q1 and F1 texts of *Henry V* [160], lists three variant readings involving the expansion *all's* and *there's* in the Folio text. While the exact relationship between these versions is uncertain, the differences may support Partridge's suggestion that Shakespeare wrote out contractions involving pronouns in full until about 1600 and relied on the actors to recognize that the words must be slurred [120, p. 63]. Perhaps a more interesting example is provided by *Troilus and Cressida,* where it appears that the first 3 pages of the Folio text were set from the quarto text; for *in't* at TLN 100 in the Folio text[7] Q reads *in it.* Farnham notes that in Othello the earlier text prints the full forms of *in's* and *to't.*

All in all it seems that one cannot be absolutely certain that contracted forms (or full forms) *always* reflect an author's preference. Expanding contractions to their full forms seems to be the safest road since direct scribal or compositorial modification of these features may have occurred. On the other hand, the number of contracted forms may be so small that the decision to expand or not to expand might not affect the results of any analysis.

Smith, in one of his studies of collocations and proportional pairs [148], made use of two versions of a play, one reflecting the original orthography and another containing full forms. I also thought it best to make two sets of counts in

---

[7] "TLN" refers to Hinman's "Through Line-Numbering" system. The TLN system, which is described in the introduction to the Norton facsimile [134], continuously numbers lines in the original text from the first line to the end of the play, with no reference to acts or scenes.

the current study. However, a computer program was developed to expand contractions and to standardize spelling variants in existing texts. Naturally in many cases forms in the computer files required special coding (for example, to recognize *has* as *he has* or *on's* as either *on his* or *on us*). The method is certainly not completely automatic. Before examining the details of these procedures, the selection of individual plays is examined.

2.3 The Shakespeare Texts Used in this Study

In choosing the texts to be used as control samples in this comparative study, a set of plays was selected to span the length of Shakespeare's career and the types of play he produced. (The second goal was not fully achieved with regard to the tragedies, since the available texts of *Hamlet, Othello* and *King Lear* present difficulties.) Originally eleven Shakespeare plays were chosen for a stylometric analysis, but an initial study of this sample determined that certain "habits" occurred less predictably than previous studies had assumed. In order to determine if this observed variation was a feature of the control sample or of the population, the initial set of eleven was expanded to twenty-four plays.. Twenty of these made up the *control set* used for determining the characteristics of Shakespeare's dramatic writing; the other four were used a *test set* in later stages of the study. These were treated as works of unknown authorship in order to validate all methods evaluated.

Several principles were observed in determining which plays were *not* to be included in the study. First, any play with serious questions of authenticity was eliminated. At one time or another, most of Shakespeare's plays have been questioned to some extent, but many of these claims were put forward by the over-zealous disintegrators of the late-nineteenth century and early twentieth century. Of the 37 plays that make up the Shakespeare canon, authorship questions still surround *Titus Andronicus,* the three *Henry VI* plays and *Pericles* (in

addition to one of the objects of this study, *Henry VIII*). Another play, *Measure for Measure,* may contain certain non-Shakespearean additions, according to one of the Shakespearean scholars who advised me in text selection.[8] Upon this advice this play was also not included in the control set. This omission should not be significant, since the final control set includes a large number of comedies (including *All's Well that Ends Well,* a play of similar style and date).

In choosing among the remaining plays, textual considerations surrounding the sources and versions of the early printed editions played an important role. In making use of early versions of the texts, I was in much the same position as an editor who has to select the copy-text upon which to base an edition of a play. In most cases I was only willing to select plays that have one authoritative source text, since otherwise I would have to bring together readings from several texts to form my own version: in effect, become an editor. In this case I would have done well to use modern critical editions and bow to the judgement of those scholars who have devoted their careers to such work. But considerations of the sources of the printed editions did affect some decisions, and an understanding of such factors is required in order to understand the discussion of the individual plays below. (Appendix A, *The Sources and Printing of Early Editions,* is provided for those readers unfamiliar with these areas of study.)

Eighteen of the thirty-six plays published in the Shakespeare First Folio exist in more than one early printed version, and the most difficult selection decisions have involved cases where the relationship between these multiple versions is not fully understood. However, for the other eighteen plays, one has no choice: these plays exist only in the versions printed in the Folio. Several of these fall into the category of suspect authorship described above. Another, *Timon of Athens,* has been the subject of wide debate to explain its inconsistencies of structure and language. While it appears that the text was printed from the author's foul papers, Greg feels that these must have been an early draft "that had never been

[8]This scholar has asked that as little as possible be said about these results until they are published.

reduced to anything like order" [43, p. 411]. The inclusion of such a poor text in the Folio may not have been intended initially by Heminges and Condell, since Hinman's bibliographic study [49] shows that *Timon* was printed where *Troilus and Cressida* was originally to have been included.

The problems involving multiple authoritative versions ruled out a number of plays, including most of Shakespeare's great tragedies. *Hamlet* is an exceptionally complicated textual problem; it occupies Jenkins for 64 pages in the critical introduction of the recent Arden edition [128]. There are three different versions of what some consider Shakespeare's greatest play. Both the authorized quarto published in 1604 and the Folio text are largely substantive, but appear to be derived from different sources. The earlier 1602 quarto was clearly produced by memorial reconstruction. While Q2 appears to stand closest to Shakespeare's papers, it leaves some unique passages obscure. The Folio text appears to contain some authentic additions not found in Q2 as well as some spurious additions. Both F1 and Q2 also owe something to their predecessor Q1, and Jenkins notes the possibility that even readings that all three agree upon could be wrong (p. 74).

King Lear and *Othello* also have "doubtful" quartos, as Hinman terms them in the introduction to the Norton facsimile of the First Folio. Sanders, in his recent edition of *Othello* [135, pp. 206–207], finds that both Q1 and F1 are derived from two distinct manuscripts of equal authority but both corrupted to some extent in transmission. The textual problem of *King Lear* involves a most "unusual" quarto that appears to have some authority. Until recently the Folio text was thought to have been printed from this quarto and corrected against another manuscript, but recently scholars have decided that it reflects a revision by Shakespeare himself.

The case of *2 Henry IV* is similar to that of the tragedies. Two slightly different versions of the 1600 Quarto exist, and the Folio contains several passages that were omitted from both of these, perhaps due to censorship or the shortening of the performance. But again, Humphries [137] notes that the Folio

text may have been set from a copy of the earlier quarto collated with another manuscript. The source of the 1609 quarto of *Troilus and Cressida* was certainly the author's papers, but the evidence suggests that another manuscript was used in addition to the quarto in preparing the Folio text. Thus for these five plays (*Hamlet, Othello, King Lear, 2 Henry IV* and *Troilus and Cressida*) one cannot simply choose an old-spelling version from among the quartos and Folio texts that corresponds to what modern editors feel best represents the "true" text. For this reason they were not included in the set of texts used in this study.

Twenty-four Shakespearean dramas were considered more suitable for the purposes of control and comparison in this study. These plays are listed in Table 2–1 with their probable dates and an indication of which early edition was used. Of the thirty-seven plays of the canon, the sample includes: eleven of the thirteen comedies; five of the ten histories; four of the ten tragedies; and three of the four romances. I decided that four of these plays were to be set aside for use as a test set and chose *Richard III, As You Like It, Antony and Cleopatra* and *The Tempest.* This set represents each of the four genres and is slightly biased towards the latter part of Shakespeare's career, when *Henry VIII* and *The Two Noble Kinsmen* were written.

Earlier I stated that texts that existed in more than one authoritative version would not be used, since I would be required to edit the two versions into a single text. This principle has been violated to some extent in seven plays. In four plays certain passages found in the chosen edition have been not included in the analysis.

1. In *Macbeth* three short passages are recognized as interpolations from Middleton's play *The Witch.* The fifty-eight Folio lines involved are: all of III.v (TLN 1428–1469); IV.i.39–43 (TLN 1566–1572); and IV.i.125–132 (TLN 1672–1680). This text of the play is remarkably short and seems to have been set from a prompt-book that had been cut due to censorship or for a special performance. What remains (other than these three witch passages) is regarded as Shakespeare's unaided work.

Play	Probable Date	Edition Used
The Comedy of Errors	c. 1589–1593	F1
Love's Labor's Lost	c. 1588–1589	Q1
Two Gentlemen of Verona	c. 1590–1594	F1
Richard III	c. 1591–1594	F1*
The Taming of the Shrew	c. 1592–1594	F1
A Midsummer Night's Dream	c. 1594–1595	Q1
Romeo and Juliet	c. 1594–1596	Q2
King John	c. 1594–1595	F1
Richard II	c. 1595–1596	Q1*
The Merchant of Venice	c. 1594–1598	Q1
1 Henry IV	c. 1596–1598	Q1*
Much Ado About Nothing	c. 1598–1599	Q1
As You Like It	c. 1598–1600	F1
Henry V	c. 1599	F1
Julius Caesar	c. 1599	F1
The Merry Wives of Windsor	c. 1597–1601	F1
Twelfth Night	c. 1600–1602	F1
All's Well that Ends Well	c. 1601–1604	F1
Macbeth	c. 1606–1607	F1
Antony and Cleopatra	c. 1606–1607	F1
Coriolanus	c. 1608	F1
Cymbeline	c. 1608–1610	F1
The Winter's Tale	c. 1610–1611	F1
The Tempest	c. 1610–1611	F1

*Text used includes passages from another edition; see text below.

Note: Date and order are taken from Bevington's edition [126, p. 72].

Table 2–1: The twenty-four Shakespeare plays used in this study

2. The quarto text of *Love's Labor's Lost* was set from the author's foul papers and includes two passages that seem to be first drafts of following lines. These lines, IV.iii.292–313 and V.ii.813–818, are certainly by the author, but he would not have included two versions of the same passage in the final version. Therefore these have been omitted from the text for the purpose of this analysis.

3. Q2 of *Romeo and Juliet* was also set from the author's foul papers and contains two passages similar to those described in *LLL*. On leaving Juliet's balcony Romeo speaks four lines that begin the Friar's speech in the next scene (II.ii.1–4). Again, Romeo's soliloquy at Juliet's tomb includes four lines that are clearly a first draft of the following thirteen (V.iii.108–120). These passages have also not been included in the analysis.

4. *Henry V* contains a number of speeches or passages in French. Together these make up a fairly large number of words. They were therefore marked in such a way that they would be ignored by the computer software. No attempt was made to mark the occasional word or short phrase of Latin, mock Welsh, *etc.* which can be found in this and other plays.

In three other plays the copy-text selected by modern editors is a particular edition *except* for one or two complete passages found in another version of the text. For these cases I have copied the more authoritative version of the passages into the edition that I have selected. The description of each case should justify these actions.

5. The 1598 Quarto *1 Henry IV* was thought to be the earliest published version of this drama until a single surviving sheet of an earlier quarto was discovered. Thus, modern editors use the 1598 Quarto (Q1) as copy-text except for the eight pages of Q0, and I have followed their example.

6. The 1597 Quarto of *Richard II* is a good text, set from the author's own papers; however, the deposition scene, in which Richard gives up the crown

to Bolingbroke, is conspicuously missing, probably due to censorship. A memorially reconstructed version appears in Q4, but the Folio publishers apparently made use of a better source for this scene. (The evidence for this conclusion is outlined by Ure [132, p. xv].) I have therefore taken the Folio text for this scene and inserted it into the appropriate place in the quarto text.

7. The relationship between the 1597 Quarto of *Richard III* and the Folio text is one of the most difficult textual problems facing an editor of Shakespeare. The Quarto text is a very unusual sort of bad quarto; it appears to have been reconstructed by the entire acting company (perhaps including Shakespeare himself) to replace the company's copy of the text, which presumably had been lost or misplaced. The Folio text seems to have been set from parts of reprints of this quarto which had been corrected against an independent authoritative manuscript. This proof-reading and correction process seems to have overlooked two sections of the Folio text (III.i.1–158 and V.iii.48 to the end), and for these Q1 is the more authoritative version and has been used in this study. For the remainder of the text, the Folio edition is used, although with many reservations. (The twenty or so lines that appear in Q1 but are omitted in F1 include one substantial (and well-known) passage: the "clock" scene, IV.iii.98–116, in which Richard refuses to reward Buckingham for his role in making Richard king. This passage has *not* been included in the version used in the control sample.)

The problems here would seem to indicate that for my purposes *Richard III* belongs in the same category as *2 Henry IV, King Lear* and *Othello*. It would certainly be desirable to make use of it in some way, since it is an early play with a very rhetorical style. So a version of the work has been put together (literally), but all evidence deriving from this text will be viewed with a certain amount of scepticism in some applications. In fact the play is one of the four texts that make up the test set and thus will not be used to establish Shakespeare's habits of composition.

In most cases, the alterations I have made in the copy-texts required no editorial judgement other than a knowledge of the relationship of the different early versions. The actions of modern editors in these matters are uniform and the substitutions or deletions are quite straightforward, so in these six texts I have created versions that differ from the quarto or Folio texts. The case of *Richard III* stretches my principles to some degree, but an awareness of the limitations of this text in the application of authorship techniques should prevent any problems. To some extent none of these texts is ideal: some may have been corrupted in transcription or in the theater, and all have possibly been subjected to alterations in the printing house. These are the versions of the texts that have survived, however, and one of the goals of this study is to determine whether or not they can be used effectively to solve certain Shakespearean authorship problems.

2.4 The Fletcher Texts Used in this Study

In many ways choosing a set of control texts for Fletcher is more difficult than for Shakespeare. First, he wrote many of his works in collaboration with another playwright, and there is little external evidence identifying plays as his unaided work. Second, Fletcher's plays proved popular until the closing of the theaters in 1642; in many cases there is strong evidence that they were altered for their revival. Finally, only one of the works that can be attributed to Fletcher alone was published in a quarto during his lifetime. Most were not printed until 1647, twenty-two years after his death, when the Beaumont and Fletcher First Folio was published. While these facts do make the selection of plays more difficult than for Shakespeare, Bowers' old-spelling critical editions are certainly a resource that can be used with some confidence.

2.4.1 Fletcher's Unaided Work

The first goal is to determine which plays are solely Fletcher's, of course. The most important study of the authorship problems in the Beaumont and Fletcher canon is Cyrus Hoy's "The Shares of Fletcher and his Collaborators in the Beaumont and Fletcher Canon" [54]. In this ambitious project Hoy set out to separate the many contributors to what has been conveniently if inaccurately called "the Beaumont and Fletcher canon" since it was first published by the actors of the King's Men. His criteria for discrimination were a dramatist's preference for: *ye* or *you*; third person singular verbs ending in *-th* such as the auxiliaries *hath* and *doth*; and the contractions *'em* for *them, i'th'* for *in the, o'th* for *on* or *of the, h'as* for *he has,* and *'s* for *his* (such as *in's*). These features were familiar tools in authorship studies before Hoy, but by observing them in all the plays of

the canon[9] he discovered certain patterns that had not been recognized before. Fourteen plays, listed in Table 2–2, stood out from the rest. The most striking feature of this set was the consistent use of *ye* throughout each play. In the other plays this form appears sporadically or not at all. In conjunction with the use of *ye,* Hoy noted the frequent use of contractions and the infrequent use of *-th* forms of third person singular verbs.

Hoy decides that this pattern reflects a single author's choice of linguistic alternatives, and he then shows that this man cannot be either of the other two major contributors to the canon. In the plays most closely associated with Beaumont (*Philaster, The Maid's Tragedy, A King and no King* and *The Knight of the Burning Pestle*), *ye* seldom or never occurs. In fifteen of Massinger's unaided plays, he avoids using contractions and only uses *ye* twice. Perhaps most importantly, for three of the fourteen plays external evidence exists linking them with John Fletcher.

Hoy's results in this matter seem to have satisfied scholars since his study first appeared; even the textual introductions to the plays of the very recent Bowers' editions refer the reader to this study in regard to questions of authorship.[10] Criticism of some of Hoy's results (by Schoenbaum [125] and Leech [69]) centers on Fletcher's collaborations; his initial findings regarding Fletcher's unaided work, on which all subsequent applications of the method are based, seem to have

---

[9]Hoy omitted *The Faithful Shepherdess* from his study because although "undoubtedly Fletcher's own, linguistically at least it has nothing in common with any other of his unaided works" [54, p. 142]. The Q1 text (probably published 1609–1610) includes commendatory verses by Jonson and Beaumont among others. Chambers [23, p. 222] and Bowers [9] agree that the presence of Beaumont's praise in this edition implies that he did not contribute to the work, despite Jonson's comment to Drummond in the winter of 1618–1619 that the play was written by the pair. The play's language is pastoral poetry, uncolloquial and somewhat archaic. Hoy states (p. 142) that it would be a great mistake to consider the work as typical of Fletcher, and indeed he finds that in this play the linguistic features used in his method do not match the pattern he finds in the the dramatist's other works. Perhaps for these reasons it *should* have been included in this examination, but other more typical plays were chosen instead.

[10]Perhaps one must add that Hoy did edit one play in each of the Bowers' Volumes IV and V.

Play	Probable Date	Genre	Edition Used
The Woman's Prize	1610–1611	Comedy	Bowers'
Bonduca	1609–1614	Tragedy	F1
Valentinian	1610–1612	Tragedy	F2
Monsieur Thomas	1610–1613	Comedy	F2
The Mad Lover	1616	Tragicomedy	–
The Chances	*c.* 1617	Comedy	Bowers'
The Loyal Subject	1618	Tragicomedy	Bowers'
The Humorous Lieutenant†	*c.* 1619	Comedy	MS†
Women Pleased	1619–23*	Tragicomedy	–
The Island Princess	1620–1621	Tragicomedy	Bowers'
The Pilgrim	1621?	Comedy	–
The Wild Goose Chase	1621?	Comedy	–
Rule a Wife and Have a Wife	1624	Comedy	–
A Wife for a Month	1624	Tragicomedy	–

*Possibly a revision of an older play; see text below.

†Version used is the manuscript *Demetrius and Enanthe*; see text below.

Note: Dates taken from textual introductions to Bowers' editions and from Bentley [12].

Table 2–2: Fourteen plays by John Fletcher

been accepted without question. This acceptance is not only due to the unusually convincing nature of Hoy's analysis but also to the fact that his results confirm the findings of previous critics, for the most part. Indeed, Bentley's attributions for these plays in *The Jacobean and Caroline Stage* [12] correspond to Hoy's (usually with a wry comment such as this one for *The Humorous Lieutenant:* "Oddly enough, none of the disintegrators has found any hand but Fletcher's in the play").

2.4.2 Existing Machine-readable Fletcher Texts

Of the fourteen plays in Table 2–2, four were already available in some form in machine-readable versions. A previous stylometric study commissioned by Metz

and carried out at Edinburgh by Morton[11] relied on *Monsieur Thomas* and *Valentinian* as control samples for Fletcher. Inspection of the photocopies from which the texts were prepared revealed that both were taken from the second Beaumont and Fletcher folio of 1679. For these two plays the F2 texts are not authoritative since they were printed directly from earlier published editions.

The Tragedy of Valentinian first appeared in the 1647 Folio. Turner [10, pp. 274–275] describes the fragile evidence which indicates that the play was set from a scribal transcript of the author's working papers which was later reworked by the author. "If this hypothesis is correct, the F1 text of *Valentinian* is essentially at two removes from Fletcher's papers, and because it is uncertain how thorough or careful his revision was, the F1 readings need careful evaluation." The editor of F2 took this as copy and supplied a number of substantial emendations and corrected some of the punctuation, but overall his alterations "seem not beyond an intelligent man working on his own or making only occasional reference to another text."

The main plot of *Monsieur Thomas* is complete farce, while the subplot exhibits all the characteristics of Fletcher's style of tragicomedy. The authoritative text of the play is the 1639 Quarto, which Gabler (the editor of the play in Bowers' series [10]) believes to have been based upon a fair copy in Fletcher's own hand. Again, the F2 text was clearly printed from the earlier edition.

It is quite unfortunate that these texts were not originally prepared from authoritative editions or from Bowers' critical editions. If I wished to make use of these existing texts, I felt that I had three choices. First, I could have simply used the F2 texts as they are. The textual introduction and critical apparatus of the Bowers' editions indicate no insertions or deletions of lines (or other drastic differences) between the two editions of either play. The second option would have been to edit both plays to agree completely with Bowers' edition. However, the differences in spellings and accidentals would make this a major task. The last

[11]The results of this study are unpublished.

choice would have been to edit each F2 text so that they reflect the substantive emendations listed in the footnotes and collation in the Bowers' edition. This final choice has the fault that the resulting text is really neither the early text or the critical edition, but something in between. This also assumes that all such differences are described in the apparatus of Bowers' editions.

After a great deal of consideration and some consultation with several editors, I decided to follow the first option and not edit the F2 texts. If a method of resolving authorship questions is so sensitive that the changes introduced in the F2 editions of these two Fletcher plays can significantly change the results, then one cannot hope that such a method will be useful, given the inescapable textual problems of sixteenth century drama. However, these two plays will not be included in the control sample used to determine the normal habits of Fletcher's writing. They make up the test set and will be used at a later stage of the study as *negative checks* to ensure that any method is producing expected results.

Two other Fletcher texts were available from sources outside Edinburgh. *Bonduca,* another tragedy, has many characteristics of tragicomedy, and Leech notes that it was clearly written "with *Cymbeline* in mind" [69, p. 163]. The title character is Queen Boadicea from Holinshed's *Chronicles;* the main character Caratach bears some resemblance to Sir Walter Raleigh, and the moving description of his sorrow at Hengo's death most probably reminded the Jacobean audience of the death of the Prince of Wales in 1612. A machine-readable version of the F1 text of *Bonduca* had been kindly provided by editors of the Cambridge Webster project.

The F1 version was the first published edition of the play, although a manuscript, written between 1625–1635 for a private patron, is also preserved. The manuscript was transcribed from the author's foul papers because the promptbook had been lost (as described in a note by the manuscript's scribe, Edward Knight, book-keeper of the King's Company), but these papers were obviously defective, since Knight was forced to summarize about 190 lines of text. The

F1 Reading	MS Reading	Location
tainted pleasures	*Cæsars* pleasures	I.i.37
hated ravisher	high sett ravisher	I.i.87
burn their mentions	barre their mentions	I.i.144
these, and Chibbals	cheese and chibbals	I.ii.89
his libertie	*has libertie*	II.ii.56
set up scales for Victories	sett vp stales for victories	III.v.79
melting envie	Eating Envy	IV.iii.166
bloody fears	bloody *Sears*	IV.iv.76

Table 2–3: Emendations adopted from MS *Bonduca* into the F1-based text

missing prompt-book reappeared in time to be used for the 1647 Beaumont and Fletcher First Folio. Collation of the manuscript with the F1 text not only reveals several misreadings and omissions in MS but also what might be considered revisions in F1. However, these changes may not be due to actual revision but to the treatment by Knight of marginal additions or alterations in the foul papers from which he worked. In any case the MS provides a distinctly inferior text (because of its lacuna, misreadings and omissions) than the Folio, except in eight readings. Several of these were probably altered by the censor, and I have followed the practice of Gabler [10] in adopting these substantive readings, which are listed in Table 2–3, into the F1 copy-text. (These variants are also good examples of the sort of small non-authorial readings that might occur in any printed Jacobean text.) While this action may resemble the rejected third option in the matter of the two derivative F2 texts, it is significantly different. For those texts I would have been incorporating readings from the authoritative text into the fabric of a derived text, while in the case of *Bonduca* the basic text is the preferred copy-text.

The final machine-readable Fletcher text that was already available is a manuscript version of *The Humorous Lieutenant*. Like *Monsieur Thomas* this is a comedy with a tragicomic side. The manuscript version of the play, which is

entitled *Demetrius and Enanthe*, was prepared by Ralph Crane for a private patron in 1625 and thus predates the first printed version of the play by twenty-two years. The MS text contains 66 lines not present in F1 and omits 80 other lines found in the later version (which mainly comprise the prologue, epilogue and a song). In addition to these differences, the two texts differ in single words on numerous occasions. Hoy, editor of the play in the Bowers series [11], discusses the differences between the two versions of the play in detail. He determines that both texts derive from Fletcher's original papers, although MS is the fuller text and suffered far less in transmission. But in spite of this he does not choose the manuscript *Demetrius and Enanthe* as his copy-text:

> Since, then, the text of the play is more fully and more faithfully preserved in MS than in F1, MS would almost inevitably serve as copy-text for the present edition were it not that to base an old-spelling text on it would be to bestow upon the edition a system of spelling and punctuations (that of the scribe Crane) that would present something of an anomaly among the dramatic texts of these volumes. All of Crane's characteristics as a scribe are on prominent display in the manuscript of *Demetrius and Enanthe*...

In this situation, considerations for a published edition do not hold true for an authorship study. Indeed, the Shakespeare control sample described earlier includes texts that were probably set from Crane transcripts (for example *The Tempest* and *The Winter's Tale*); these also contain features characteristic of his hand (although not to such an extent as *Demetrius and Enanthe*). To be effective an authorship method must be relatively immune to changes in accidentals introduced by scribes or compositors. Therefore there is no reason why the machine-readable version of the manuscript cannot be used with confidence in this study.

2.4.3 Fletcher Texts Prepared for this Study

Four Fletcher plays were thus available for use in this study, but two of these were taken from derived texts and could not be used fully to determine the playwright's habits. More of Fletcher's work was required, and I therefore chose four

plays from the remaining works of Table 2–2 to be typed in by the Data Preparation Group of the Edinburgh Regional Computing Centre. For the reasons outlined earlier in this chapter the critical old-spelling editions published in the Bowers series were used. This decision reduced the number of plays from which to choose, since four of the remaining ten plays (*The Pilgrim*, *The Wild Goose Chase*, *Rule a Wife and Have a Wife* and *A Wife for a Month*) had not yet been edited and published by Bowers' team when I began this study.[12] These plays were written late in Fletcher's career, and therefore are further from the days when he may have collaborated with Shakespeare. Because of the time and effort involved in preparing and proof-reading a play, it was decided that four of the six plays would be used.

One of the plays to be excluded was *The Mad Lover*. Turner [11] concludes that the printer's copy for the authoritative F1 text was Fletcher's foul papers "partially worked over at least once by another agent, who may have tampered with the dialogue in undetectable places", but his "impression, however, is that the lines were left pretty much as Fletcher wrote them." Since the editor states that textual revision is a more-than-usual possibility, *The Mad Lover* was not chosen for the Fletcher control sample.

The other play to be excluded is *Women Pleased*. A problem involving the play's date and the possibility of revision stems from an apparent reference to this play in Shakespeare's *The Taming of the Shrew*, in which a character in the Induction alludes to a clown and an incident in Fletcher's play or an Elizabethan version of it. (Naturally the lines in *The Shrew* could be a late insertion into the text, first published in the 1623 Shakespeare Folio, in which case no earlier version of Fletcher's play need be postulated. However, in the speech containing the allusion an actor's name has been set accidentally, and there is no record of this actor after 1604.) Bentley [12, p. 432] concludes that the theory of Fletcher's

---

[12]The four plays and *Wit Without Money* are included in Volume VI of the series, which was published in 1985.

revision of another author's work is not unlikely, since the play is "curiously ill-constructed" with elements "too old-fashioned for the latter part of Fletcher's career." Gabler (who edited the play in Bowers' series) states that although no one has isolated any revisions within the text of *Women Pleased,* the theory of revision regarding the play cannot be conclusively disproved [10, p. 445]. All one can conclude from the facts is that "the text and (lost) holograph manuscript of John Fletcher's *Women Pleas'd* in its extant version must be considered as ~~of~~ a play in the repertory of the King's Men in 1619–23." This uncertain state of affairs was seen as sufficient grounds for leaving this play out of the Fletcher sample.

The four remaining plays, *The Woman's Prize, The Chances, The Loyal Subject* and *The Island Princess,* were prepared from the Bowers editions. The texts of these plays are not entirely without their problems. *The Women's Prize* was probably written early in Fletcher's career and first acted around 1611. This vigorous comedy is clearly a reply to Shakespeare's *The Taming of the Shrew;* in it Fletcher uses the character names Petruchio, Bianca and Tranio. But typically Fletcher inverts the action, and the women conquer the men quite convincingly. Indeed, Appleton judges that "the Petruchio of Shakespeare's play has suffered psychic emasculation" in his search for a second wife [2].

There are two references to the play's revival in 1633; it was presented at court on 28 November, and the previous month it was the subject of a well-documented incident (described by Bowers [10, pp. 3–5]) in which the Master of the Revels, Sir Henry Herbert, called the prompt-book in for censorship. Receiving the book on a Friday afternoon, he returned it the next Monday "purgd of oaths, prophaness, and ribaldrye." Herbert remarks that this begins a new policy in which revived plays must be resubmitted for his approval (and therefore another licensing fee paid to him). Critics agree that the text printed in the 1647 Folio derived from this revised prompt-book. However, a manuscript exists that appears to have been prepared from an uncensored prompt-book. Although some censorship may be indicated by the existence of softened oaths in MS, these may have been

due to a conscientious scribe (probably Knight again) anticipating censorship or adjusting the text for a particular patron. Bowers determines that the F1 text is closer to the author's papers. He thus uses it for the copy-text of his edition, but he also uses the manuscript to correct compositorial errors and to determine the extent of the revisions due to the 1633 censorship.

There are several reasons to choose *The Woman's Prize* rather than *Women Pleased,* although both may have been subject to revision. Revisions due to censorship would probably be far less extensive than might be found when an older play was rewritten (as might be the case in *Women Pleased*). Moreover, the existence of a second text of some authority allows one to better evaluate such changes in *The Woman's Prize*. Finally, this play was written much nearer the time when *Henry VIII* and *The Two Noble Kinsmen* were composed. The only precaution taken is the exclusion of the prologue (which refers to Fletcher in the third person) and the epilogue from the analysis.

The Chances, another comedy, was very successful after the Restoration, although Appleton quickly dismisses it: "Of *The Chances* little need be said." Although no external evidence exists about performances before 1630, two items of internal evidence point to a revival soon after the author's death. The prologue mentions Fletcher's "lovd memorie," and a short passage appears to allude to political events of early 1627. Therefore both the prologue and this passage (starting in III.I.4 with "Yee shall..." and ending in verse 10 with "...so well neither.") have been removed from the text used in this study.

The tragicomedy *The Loyal Subject* may make a political statement about Sir Walter Raleigh and James I in its description of the mistreatment of the old general Archas by the young Duke. Bowers determines that the relatively clean F1 text was set from prompt copy [11], but again the prologue refers to Fletcher's death, "our now widdowed stage In vain lamenting." This play is one of the three which external evidence links with Fletcher alone. Herbert, the Master of Revels, has noted in November 1633 that the play, "an ould booke of Fletchers," had originally been licensed in 1618. He also speaks of "some

reformations" that he made for the 1633 revival, which Bentley [12, p. 372] feels means slight alterations by the company rather than his censorship. Bowers does not discuss this question in his textual introduction to the play. Again, the prologue and epilogue were not included in the analysis.

The last of the texts from Bowers' edition that has been used is the tragicomedy *The Island Princess.* Like *The Chances,* this play enjoyed a huge success after the Restoration, although most probably always in an altered version. Williams [11] discusses in detail the features of the only authoritative text (from the 1647 Folio), but does not even suggest the nature of the copy behind this text. None of the features described seem to suggest foul papers, however.

Eight plays by Fletcher alone are thus available for purposes of comparison and control in this authorship study. Two of these, *Monsieur Thomas* and *Valentinian,* are taken not from the most authoritative versions but from the second Beaumont and Fletcher folio, and their value is thus slightly tarnished. One of the remaining six, *Demetrius and Enanthe,* is a manuscript version. Another, *Bonduca,* is a copy of the authoritative Folio text with eight substantial emendations from MS. The remaining four texts (*The Woman's Prize, The Chances, The Loyal Subject* and *The Island Princess*) were prepared from the most recent old-spelling, critical editions. Four of the eight plays are classed as comedies, two as tragicomedies and two as tragedies, but one must remember that Fletcher often included a tragicomic element in both comedy and tragedy. The possibility of revision due to revival or censorship seems much more strong in these texts than in Shakespeare's, but this may simply reflect our ignorance of the stage history of the older dramatist's plays. Again it must be stressed that the possibility of small non-authorial alterations in a text is an inescapable part of the problem in almost every attribution study of English Renaissance dramas, and therefore one had best devise methods as insensitive as possible to small, local interpolations or revisions.

2.5 Computer Processing of Old-Spelling Texts

Those who have never undertaken a computer-assisted textual study might imagine that all of the tedious work would fall upon the machine's broad shoulders. But a large proportion of the time spent on this project was spent proof-reading texts or marking certain textual features in computer files. (Perhaps this is less surprising when one considers that the thirty-four plays used in the study contain almost 750,000 words.) The use of a machine for searching and counting does not release the researcher from the responsibility of closely examining the texts to ensure that the computer software is able to produce accurate results.

While this applies to any computer-based study, characteristics of these dramatic texts in original orthography create particular headaches. Spelling variations and homonyms present in these texts make it difficult to use computers to count all forms of common words such as *do* or *I*. Automatic recognition of the elements of compound contractions is another difficult task. Smith, in the only previous stylometric study that has considered the extent of contractions (which will be described in Section 3.3.3), prepared two versions of each text. The procedures outlined in this chapter can be used to produce *expanded texts* from existing text files if certain word forms are marked before processing.

2.5.1 Proof-reading, Data Format and Light Editing

In the early stages of the research proof-reading was a major chore. Naturally the newly prepared Fletcher texts required detailed examination. In addition, the Shakespeare quarto texts received from Oxford appeared to be early versions of files that had not been completely corrected after the initial data entry process. (Contacts in Oxford believe that the final versions used for the published concordances have been mistakenly destroyed.) Speech prefixes and stage directions

were sometimes incorrectly marked and occasionally several lines were missing from the computer files. Thus the eight quarto texts obtained from Oxford that were used in this study required considerable proof-reading (about one month's work) against facsimile reproductions of the original quartos [138].

Most projects in literary computing require a text to be formated and marked in a special manner; the current study is no exception. One important requirement is the ability to distinguish the speeches in the play from other information printed in the text. Speech prefixes, act and scene headings and stage directions are not part of a playwright's creative composition, although they may provide valuable evidence in determining the source behind the text or the manner in which an early edition was printed. Stage directions were often added or modified in the prompt-book during production, and the final appearance of headings and speech prefixes was usually due to the compositors. However, these textual features should not be deleted from the computer file, but marked with a system of *reference identifiers* so that a program can retrieve this information for the user.

The marking system used in this study is sometimes referred to as "COCOA-format" references, after the concordance-generation program developed at the Atlas Computer Laboratory. This system has been adopted by COCOA's successors: CONCORD, a program developed here at Edinburgh and used in some stages of this study, and the more recent (and more widely available) Oxford Concordance Program (OCP). While reference identifiers sometimes set off words that are part of the original edition,they are also used to include information in the computer file that is not part of the original text. In either case the reference is composed of a one-letter identifier followed by a string of words, with a pair of angle-brackets surrounding the entire reference. Figure 2–1 shows the beginning of the computer file containing *Macbeth* and demonstrates the encoding of reference information.

The texts prepared at Oxford made use of this system, and each play contained identifiers for compositors, page signatures and line numbers according to

```
<T Mac><P 116><C A>+
<I 1.1>+
<L 1> <Z {Actus Primus. Scoena Prima}.>
<L 2> <D {Thunder and Lightning. Enter three Witches}.>
<L 3> <S 1.>When shall we three meet againe?
<L 4> In Thunder, Lightning, or in Raine?
<L 5> <S 2.>When the Hurley-burley's done,
<L 6> When the Battaile's lost, and wonne.
<L 7> <S 3.>That will be ere the set of Sunne.
<L 8> <S 1.>Where the place?
<L 9> <S 2.>Vpon the Heath.
<L 10> <S 3.>There to meet with {Macbeth}.
<L 11> <S 1.>I come, {Gray-Malkin}.
<L 12> <S {All}.>{Padock} calls anon: faire is foule, and foule is faire,
<L 13> Houer through the fogge and filthie ayre.<D {Exeunt}.>
<I 1.2>+
<L 14> <Z {Scena Secunda}.>
<L 15> <D {Alarum within. Enter King Malcome, Donal%baine},>
<L 16> <D {Lenox, with attendants, meeting}>
<L 17> <D {a bleeding Captaine}.>
<L 18> <S {King}.>What bloody man is that? he can report,
<L 19> As seemeth by his plight, of the Reuolt
<L 20> The newest state.
```

Figure 2–1: The beginning of the computer file containing *Macbeth*

Hinman's "Through Line-Numbering" (TLN) system. I also added TLN numbers to the computer files containing the Shakespeare quartos and the Fletcher texts. The TLN system is especially useful in computer-based studies where simple search programs, such as UNIX's **grep**, often just print a given line with no other information. Identification is simplified when a unique line number is actually attached to every line in a computer file.

Scene headings and stage directions were also set off as reference identifiers. I added act and scene markings: for Shakespeare using the scene division of the Bevington edition (which is based on the 19th century Globe text) and for Fletcher the division used in the Bowers series. (When plays were processed by act, prologues were considered to be part of the following act and a play's epilogue part of Act 5.) The induction in *The Taming of the Shrew* was treated as a separate act.

A number of features were already marked in the Oxford quarto and Folio texts. Text in italic font was surrounded by curly brackets. "Turn-overs" and lines of text that filled the width of the Folio column were also marked (since spelling may have been adjusted in order not to break the line). A number of printer's contractions were also indicated. Often when a compositor was running out of room in a line he would delete the letter M or N in a word, marking the contraction with a *tilde* above the preceding letter:

AYL 2580: *Ros.* Patience once more, whiles our cõpact is vrg'd...

The Oxford texts included the deleted letter and marked it with a dollar sign. Printers' contractions such as $\overset{e}{y}$ and $\overset{h}{w}$ were typed simply as "y" and "w" in the Oxford files.

A certain amount of light editing was done on the basic texts. All printer's contractions and abbreviations (such as *Lo:* for *Lord*) were expanded to their full forms. In addition, words broken across a line were reunited to make things easier for both man and machine. The hyphen that marked these forms has been converted to a percent sign, as in the stage direction of line 15 in Figure 2–1. This code and the "vertical bar" symbol |, which marks a turn-over occurring between words, would allow the original line breaks to be recovered if there was ever any need.

The printing conventions for the letter pairs *i–j* and *u–v* were undergoing change during the period between 1597 and 1679, the earliest and latest publication dates of the texts used in this study (corresponding to *Richard II* and *Valentinian*). Barber, in *Early Modern English* [7], describes how the members of each pair were originally just alternative ways of writing the same letter, but in the 16th and early 17th centuries printers conventionally used *v* as the first letter of the word and *u* in every other position. The letter *j* was only used in the combination *ij*, as in *diversifijng.* The modern convention of using *u* and *i* to represent vowels became the standard practice about 1630.

The 24 Shakespeare texts used in this study conform almost perfectly to the older conventions, although things were beginning to change when the 1623

Folio was printed: *lov'd* and *lou'd* are printed alongside one another in one line of *As You Like It* (TLN 1854). *The Two Noble Kinsmen* and the Fletcher texts show some variation of usage. For my purposes it is necessary to recognize two graphical forms as the same word, so words in the new convention were converted by a program (called NEW2OLD) to the older conventions. It would have been preferable to convert the older forms to the modern conventions, but this was impossible to do automatically, whereas the rules for the older spelling conventions are easily programmed. (However, the program fails to standardize a few compounds which appear in the texts, such as *thereupon* and *therevpon.*)

2.5.2 Word Division

Recognizing the divisions between words is not as large a problem in Early Modern English dramatic texts as it is for texts from the Middle English period. Problems in these plays center on hyphenated forms and some pronoun forms. Some of the changes described in this section may emend features that reflect authorial intent. However, standardization of the different uses of these forms in the 34 texts was deemed necessary for recognition of some common words.

The use of hyphens to join words varies tremendously and often reflects the scribe rather than the author. Ralph Crane (thought to have prepared transcripts of several Shakespeare plays for the Folio publication) had a remarkable penchant for hyphenation, which is abundantly evident in his manuscript of *Demetrius and Enanthe*:

TLN 520: say you find such a-One...

TLN 3006: and poore-beleeuing I, became his Seruant...

Hyphenation is used by a writer at times to indicate the manner in which an actor should speak the lines, as this breathless outburst in *The Comedy of Errors* (thought to have been printed from Shakespeare's foul papers) clearly demonstrates:

> ...Along with them
> They brought one *Pinch,* a hungry leane-fac'd Villaine;
> A meere Anatomie, a Mountebanke,
> A thred-bare Iugler, and a Fortune-teller,
> A needy-hollow-ey'd-sharpe-looking-wretch;
> A liuing dead man.

Every hyphenated form in all the plays was examined and altered if it significantly differed from modern usage. In each instance I consulted the Bevington, Riverside and Arden modern editions and used my own judgement when these disagreed. The simplest way of separating these forms would have been to insert a space character, but it seemed prudent to maintain a distinction between white space found in the original text and space added later. Thus the backslash character \ was inserted and then treated as a space in subsequent computer processing. For example, the mouthful from *The Comedy of Errors* was altered to:

```
needy\-\hollow-ey'd\-\sharpe\-\looking\-\wretch
```

and the two examples from *Demetrius and Enanthe* were similarly separated. A hyphen on its own is easily ignored by word-counting software.

No separate words that are hyphenated in modern editions were joined. However certain modern words were represented by two words in the Jacobean period. Again the usage varies within the thirty-four texts used, and some form of normalization was considered necessary where the words involved are common function words or pronouns. Forms of reflexive and emphatic pronouns (*yourself, myself,* etc.) occur as one word or two in the same text. Since I was interested in counting these as a group and distinguishing them from possessive adjectives, these were joined together. Examples of *tmesis,* where the two parts are separated by another word ("my crying self"), were not altered. In most instances the words *today, tomorrow* and *tonight* occur as two words. This is a relic of an obsolete use of the preposition *to* and the words were also joined to correspond to the modern usage. For similar reasons the pair *an other* was joined.

A number of problems arise when considering word division and contracted

forms. Most of these were solved using the techniques outlined in the next two sections, but another special character, the forward slash /, is used in contracted forms like `th'/allusion`, `th/other` and `on/'em`. The backslash is always treated as a space character, but the forward slashes in these forms were processed differently at different stages of the analysis. This allowed these forms to be either counted as single words or expanded to their full forms.

2.5.3 Recognizing Homonyms and Variant Spellings

The advantage of using a computer in a study such as this is that it should be able to count common words in a large body of text quickly and accurately. But as demonstrated in Section 2.2.1 functions words in Jacobean texts can exist in a number of variant forms. Homonyms are also a thorny problem. In some analyses one might want to distinguish occurrences of the modal verbs *will* or *might* from the noun forms. The problem is compounded in 16th century dramatic texts because of spelling conventions and the attempt to indicate how the words should be spoken on stage. For example, *a* is usually the indefinite article but commonly represents an unstressed form of *he*. It also commonly stands for *ah* or *have*, and occasionally occurs in a prepositional form (now obsolete for the most part): "to be a weary of thee" and "I am a horsebacke," for example. Less commonly *a* represents *on*, *of* or *in*. The interjection *aye* was often spelled like the first-person singular pronoun *I*. Finally, the fact that playwrights, scribes or compositors often did not distinguish clearly between *to* and *too* or between *of* and *off* also presents a problem. Many of these problems would be resolved by using modern editions, but clearly certain homonyms and weakened forms would still need to be recognized.

Those who prepared the Oxford texts recognized many of these problems and attached a "hash" sign (#) to the beginning of some words. This marking indicated "this is somehow distinguished from the ordinary form" but did not supply any further details. For example, there was no way of automatically recognizing that an occurrence of `#a` represented *he* or *have*. A more complex

```
a'th=o'th'
a'th'=o'th'
al=all
alls=all's
alreadie=already
an#1=and
an#2=on
an#4=Anne
an#5=and
an'#5=and
an't=and't#1
and#1=and
angrie=angry
anie=any
on#1=one
one#1=on
ons#2=on's#2
ont=on't#1
or#1=our
oth#1=oath
oth'=o'th'
our#1=ours
ourselfe=ourself
```

Figure 2–2: Excerpts from spelling variants translation list

system of coding was required if one wished to include the counts of these forms with the occurrences of the more usual forms of the word. A system of *hash suffixes* was developed: the hash sign was moved to the end of the word and followed by a number identifying its "translation." To discover what words were marked in the Oxford texts, a text editor was used to find and print each line containing a hash sign. These were then sorted by the marked word, and after examining all 3187 occurrences I determined what translations were required and worked out a coding scheme. The seven Fletcher texts that did not originate from Oxford had to be examined for occurrences of word forms requiring marking. The Oxford texts were also examined to make sure that they were consistently marked for some common or important forms; some mistakes were found and corrected.

To achieve some standardization of spelling for frequent words, a program called REPLACE was developed that used a translation list to replace common spelling variants and certain words marked with hash suffixes. Forms recognized

in past studies as markers of Shakespeare's or Fletcher's preferred usage (for example, *ye, 'em, hath* and *doth*) were not altered by any of the techniques presented in this chapter. Figure 2–2 contains excerpts from the translation list, and Appendix C contains the entire list. Some attempt was made to correct more general classes of variants.[13] Variants for the most common forms were compiled from my own knowledge and proof-reading experience. When alphabetical word lists were produced by this process, these were examined and new variants noted and added to the translation list. The final word counts reflect several stages of this process.

In compiling a replacement list, one must make decisions about how some common word-forms will be counted. Certain occurrences of common words were marked to distinguish homonyms or differing grammatical function. Usually the word and the hash suffix were replaced by the translation, but sometimes words marked with hash suffixes were preserved in the standardized text files produced by REPLACE. For example, in order to distinguish modal verbs, the standardized files (and the word lists produced from them) contain occurrences of `will` and `will#1`. Likewise the second person singular form of *be* is *art*, but `Art#1` remains in the computer files to represent the noun.

On the other hand, my basic texts preserved the Oxford markings of uses of *and* and *an* as a subordinator meaning *if*.[14] I decided to disregard this functional distinction and simply count all such occurrences as *and*. The translation list

[13]For example, all words ending in *-ie* were listed and examined. If these forms would always be spelled with *-y* in modern English, they were added to the original translation list. Unfortunately the computer environment in which REPLACE was used imposed a limit on the number of variants that could be recognized in a single run. Therefore a number of words ending in *-ie* were deleted from the list. The complete list given in Appendix C is rather a hodge-podge but does include the variants for common and important function words.

[14]Part C.1 of the *OED* definition of *and*. For example:

> *CE* 259: Nay, and you will not sir, Ile take my heeles...

This meaning was sometimes strengthened by the explicit use of a following "if:" *and if* or *an if*.

thus includes entries which replace occurrences of the coded forms and#1 and an#1 with the ordinary form and.

In Figure 2–2 the form an#5 is also changed to and. This entry is included for consistency with an'#5, which is the code used in rare contractions of *and* like "eleuen an'aday." Forms like this are represented in the computer file using both the hash code and the forward slash character, which can be treated as a space or ignored: eleuen an'#5/a/day.

The use of hash suffixes with *and* demonstrates another important benefit of the coding and replacement strategy adopted here: if for some reason in the future I change my mind and decide that it would be desirable to recognize the different grammatical distinctions of the word *and*, the original data files still preserve this information. I can simply alter the translation lists and produce another version of the standardized files.

Other words with hash suffixes were not replaced in the new versions of the files. Single letters used in abbreviations were marked with #0 (hash-zero) and left in the text. For example, in *Twelfth Night* Malvolio's attempt to decipher the mysterious letter is represented as follows:

```
<L 1120><S {Mal}.> {M#0.0#0.A#0.I#0}. doth sway my life...
```

Intuitively, the "hash-zero" code means that the graphical form represents itself rather than having any independent meaning. This code is also used in cases where a character echos a word or phrase, like in this exchange in *Anthony and Cleopatra*:

```
<L 1087><S {Mes}.> But yet Madam.
<L 1088><S {Cleo}.> I do not like but#0 yet#0, it does alay
<L 1089>The good precedence, fie vpon but#0 yet#0,
<L 1090>But#0 yet#0 is as a Iaylor to bring foorth
<L 1091>Some monstrous Malefactor.
```

In a modern edition, such occurrences would be surrounded with quotation marks. The decision not to count these as occurrences of *but* and *yet* was mine.

These echos and deliberate repetitions are very "non-random" and relatively unusual. It seemed wisest to note the distinction; when analyzing the data at a later stage, counts for the hash-zero words could easily be combined with those for the usual forms. (But this was *not* done in my study of word rates described in Chapter 5.)

At this stage I had two versions of the text files: the basic text files containing the variant spellings and certain homonyms marked with hash suffixes, and a set of texts with normalized forms for some common words.

2.5.4 Expanding Contractions

In Section 2.2.2 different forms of contraction in 16th and 17th century texts were discussed. Recognizing that in some cases the author's intentions are lost, it seems desirable to be able to count contracted forms of common words. Expanding compound contractions is not such a simple problem as standardizing variant spellings and recognizing homonyms. For most cases the same basic strategy can be used: provide a list of contractions and their expansions and use a program to do the replacements. The spelling variants replacement program REPLACE has an option to expand contractions given a list of contractions and their full forms. The expansion list used was drawn up from the examination of the hash markings in the original Oxford texts and my own proof-reading experience, then checked against the many lists of contracted forms found in Partridge's book [120]. In addition, a complete word list for all the plays used in this study was searched, and every word containing an apostrophe examined. The complete expansion list is found in Appendix D.

This simple strategy based around a translation list was not used for words ending in apostrophe-*s* or apostrophe-*t*. These endings occur with so many words that explicitly listing the possibilities is not practical. Replacing all occurrences of these endings would be disastrous. Apostrophe-*s* usually indicates enclitic *is* but can often stand for *his* or *us*. Apostrophe-*t* often indicates syncope in the preterite or past-participle endings of weak verbs, such as *banish't*, but otherwise

```
he's ⟹ he is
Banquo's#1 ⟹ Banquo's
on's#2 ⟹ on his
on's#3 ⟹ on vs
within's#4 ⟹ within this
he's#5 ⟹ he has

to't#1 ⟹ to it
too#1 ⟹ to        (from the spelling variant list)
too#1't#1 ⟹ to it
```

Figure 2–3: Some results of contraction coding and expansion

represents enclitic *it*. Hash suffixes are used to identify the correct expansion in these two cases.

Of course this means examining each occurrence of a word ending in apostrophe-*t* or -*s*. Fortunately for this purpose, the use of apostrophe-*s* to indicate possessive genitive is rare in Shakespeare's and Fletcher's texts. Also, the other modern contraction associated with this ending (enclitic *has*) was extremely rare.[15] All words ending in apostrophe-*s* were examined and marked according to the coding scheme shown in Figure 2–3 (which also shows the expansion for some encoded words). All words ending in apostrophe-*t* for enclitic *it* were coded #1. The program REPLACE recognizes these hash suffixes and expands them appropriately. The program also recursively examines each element of an expanded contraction and replaces any spelling variants (for example, see the expansion of *too't* in Figure 2–3).

---

[15]Barber [7] outlines the complex rules for determining when *has* or *be* is used as the auxiliary in the perfect tenses. Apostrophe-*s* contractions in these situations were individually examined, and only one was found that might be a contracted form of *has*. In Fletcher's *Monsieur Thomas* the hero's mischievous plans are succeeding:

TLN 2489: `This Nunnery's#5 faln so pat too, to my figure...`

Variant Translations:	Contraction Expansions:
'had=had	'thad=it had
'has=has	'thas=it has
'has#1=h'as#1	h'ad=he had
'haue=haue	ha't=haue it
't'had='thad	i'ue=i haue
a#7=haue	t'has=it has
a'#7=haue	t'haue=to haue
h'as=has	t'had=it had
ha#1=haue	th'haue=they haue
ha'=haue	thou'st=thou hast
ha'#1=haue	w'haue=we haue
ha's=has	y'aue=ye haue
ha'st=hast	y'haue=ye haue
has't=hast	y'had=ye had
hast#1=haste	ye'ue=ye haue
	you'ue=you haue
	'has#1=he has
	h'as#1=he has
	ha's#9=he has
	has#1=he has
	has't#2=hast thou
	hath'#1=he hath
	hath#1=he hath

Figure 2–4: Translations and expansions for forms of *have*

The many elided and contracted forms of *has, had* and *hast* required particular care both in setting up the variant translations and contraction expansions and in marking the forms in the texts. Figure 2–4 lists entries for forms of *have* from both the translation and expansion lists. After careful examination of a text, this procedure can successfully convert the text file to a version in which the basic lexeme can be recognized from the various forms.

Admittedly this coding and replacement process is not completely automatic, but computer software that implements regular expression searches (like UNIX's search utility **grep** or a powerful text editor like **emacs**) can facilitate the initial

coding process immensely.[16] Nevertheless the pre-editing of the 34 plays used in this study was extremely time consuming.

A version of each text, containing expanded contractions, was prepared from the version with standardized spellings. At this point three versions of any given text were available: the basic version marked with the hash suffix codes; a version with standardized spellings of common words; and a version with standardized spellings and expanded contractions. These third versions will be referred to as the *expanded texts.* Where I thought contraction might influence the results of a given authorship method counts were made from both the second and third versions in order to observe the changes in the counts.

A passage from Crane's manuscript of *Demetrius and Enanthe* in all three forms is shown in Figure 2–5. This is by far the most complex and difficult text to process satisfactorily using computer software. The transformation of spelling variants and contractions can be traced in the three versions. Occurrences of *there's* in line 629 and *let's* in line 639 demonstrate the hash coding for contractions, and the occurrence of *wilbe* in line 640 shows how a spelling variant of a contracted form is modified in two steps.

Examination of the different versions of the texts indicates that this approach to variants, homonyms and contracted forms was very successful. Certainly some occurrences of common words have slipped through the system and have not been counted as they should have been. These procedures cannot cope with all the problems of 17th century orthography. For example, some modern editors (but not all) see a contraction of *is* in I.i.101 of *A Midsummer Night's Dream:*

My fortunes euery way as fairely rankt.....

Forms such as this one, where contraction cannot be recognized from the orthography, may occur in the 34 plays processed in this study. The example from *MND* is actually ambiguous, but in other cases the early editions require emendation. As it stands, lines 26–27 of IV.iii in *The Two Noble Kinsmen* will have

[16]Examples of regular expression searches will be given in Section 2.5.5.

```
<L 629> <S {Tim}.> there's no yong Wench, let her be a Sainct,
<L 630> (vnles she liue i'th' Center) but she finds her;
<L 631> and euery waie prepares addresses to her;
<L 632> yf my Wiffe would haue followed her course ({Carinthus})
<L 633> (her lucky course) I had the day before him:
<L 634> O, what might I haue byn, by this time (Brother)
<L 635> But she (forsooth) when I put theis things to her
<L 636> (theis thinges of honest Thrift) groanes, O my conscience:
<L 637> the load vpon my Conscience: When, to make vs Cuckolds,
<L 638> they haue no more burthen, then a brood Goose (Brother)
<L 639> But let's#3 doe what we can: though this wench faile vs,
<L 640> an_other, of a new way, wilbe lookd at:
<L 641> Come, let's#3 abroad; and beate our braines: Time may
<L 642> (for all his wisdome) yet giue vs a day. = <D {Exeunt}.>
```

Version 1: The basic text

```
<L 629> <S {Tim}.> there's no young Wench, let her be a Sainct,
<L 630> (vnless she liue i'th' Center) but she finds her;
<L 631> and euery waie prepares addresses to her;
<L 632> if my Wiffe would haue followed her course (Carinthus)
<L 633> (her lucky course) I had the day before him:
<L 634> O, what might I haue been, by this time (Brother)
<L 635> But she (forsooth) when I put these things to her
<L 636> (these thinges of honest Thrift) groanes, O my conscience:
<L 637> the load vpon my Conscience: When, to make vs Cuckolds,
<L 638> they haue no more burthen, then a brood Goose (Brother)
<L 639> But let's#3 do what we can: though this wench faile vs,
<L 640> an_other, of a new way, willbe lookd at:
<L 641> Come, let's#3 abroad; and beate our braines: Time may
<L 642> (for all his wisdome) yet giue vs a day. = <D {Exeunt}.>
```

Version 2: Spellings standardized

```
<L 629> <S {Tim}.> there is no young Wench, let her be a Sainct,
<L 630> (vnless she liue in the Center) but she finds her;
<L 631> and euery waie prepares addresses to her;
<L 632> if my Wiffe would haue followed her course (Carinthus)
<L 633> (her lucky course) I had the day before him:
<L 634> O, what might I haue been, by this time (Brother)
<L 635> But she (forsooth) when I put these things to her
<L 636> (these thinges of honest Thrift) groanes, O my conscience:
<L 637> the load vpon my Conscience: When, to make vs Cuckolds,
<L 638> they haue no more burthen, then a brood Goose (Brother)
<L 639> But let vs do what we can: though this wench faile vs,
<L 640> an_other, of a new way, will be lookd at:
<L 641> Come, let vs abroad; and beate our braines: Time may
<L 642> (for all his wisdome) yet giue vs a day. = <D {Exeunt}.>
```

Version 3: Spellings standardized and contractions expanded

Figure 2–5: Three versions of the text of *Demetrius and Enanthe*

an extra *the* after expansion:

```
tis a sore life they haue i'th Th/other place...
```

These problems are unavoidable without critical old-spelling editions, but fortunately such occurrences are relatively infrequent.

2.5.5 Software Used to Count Textual Features

At this point, the computer software used in this study to count textual features will be described, but not in extremely thorough detail. The best way of counting textual features in any study depends to a large extent on the local computing environment, so my experience with particular tools may not be useful to other researchers. Several sophisticated commercial software products that can count textual features, such as the Oxford Concordance Program and Brigham Young University's CONCORDANCE, are now available in many academic institutions. If counting is straightforward, then the efficient storage and retrieval of counts and associated data is perhaps of more interest. This section will also include descriptions of my experiences in this regard.

Chapter 4 will describe an analysis of collocations, which can be thought of as word pairs or patterns. Initially collocations were hand-counted from "keyword in context" (KWIC) concordances generated by the program CONCORD. This program was written at the University of Edinburgh in the language IMP and runs only under the EMAS operating systems. (CONCORD is similar to the more widely known Oxford Concord Package but runs more efficiently than OCP under EMAS.) These early counts were then transferred to data files on the Computer Science department's VAX VMS system, where they were processed with the relational database system VMS DATATRIEVE. The database program was used to find, sort and arrange the counts in such a manner that the output reports could be fed directly into programs written to perform χ^2 and other statistical tests. Statistics for each test and each text sample were then fed back into DATATRIEVE to facilitate analysis of the results.

Later a program called COLLOC was written to generate counts automatically for single words and collocations. This program was used to count collocations by acts in all the texts (both the unexpanded and expanded versions). The output from this program could be automatically converted to the format required by DATATRIEVE, using scripts of text-editor commands. Overall this process proved much more efficient and error-free than the earlier method of counting from printed concordances. Several discrepancies were found and traced to mistakes in the initial hand-counts or to errors in typing these counts into computer files. Once again the empirical result (reported for example by Mosteller and Wallace [113, p. 7]) was proved true: "people cannot count, at least not very high."

Counting words is much more straightforward than recognizing collocations. In fact, in the UNIX operating system a sophisticated program is not required. A number of standard UNIX utilities can be called sequentially to count all words in a text file. To make things easier for the user, these commands can be put into a file to make a *shell-script,* which can then be called like an ordinary system command. The first thing this shell-script must do is to remove everything that is not part of a word, including COCOA-format references and punctuation. The stream-editor `sed` was used for this purpose (but any other "programmable" editor could be used). At this stage word boundaries must be considered; the backslash character \ is converted to a space, and the contraction marker / either is or is not converted to a space (depending or whether contractions are being expanded).

Next the shell-script converts all uppercase letters to lowercase, using the command `tr`. The utility `awk` is then used to count words. `awk` is really a very simple yet powerful programming language, ideally suited for handling data arranged in columns. If a file contains nothing but words separated by spaces, the following program produces a list of every word type, plus its count and rate per thousand (program comments follow the # character):

```
awk '{ i = 1
       while (i <= NF) { # For every word in the line
         ct[$i]++        # Increment counter for word
         ++i
         total = total + 1} # Keep count of total number
       }
# At end of file, print all counts
    END {for (x in ct)
      print x, ct[x], 1000*ct[x]/total }'
```


```
awk '{ i = 1
       while (i <= NF) { # For every word in the line
         ct[$i]++        # Increment counter for word
         ++i
         total = total + 1} # Keep count of total number
       }
# At end of file, print all counts
    END {for (x in ct)
      print x, ct[x], 1000*ct[x]/total }'
```


Finally, the output is sorted by the first column (the word). An extract from a resulting alphabetic word-count list is shown in Figure 2–6. Complete word-count lists were produced for each play (both expanded and unexpanded texts), and lists for the Shakespeare and Fletcher control texts were combined to form a complete set of word counts for each author.

The simplicity of the structure of word-count files like Figure 2–6 is a great advantage when using an operating system like UNIX. A number of standard utilities such as **sort** and awk operate with data arranged into columns and are very efficient. The **grep** family of search commands (which quickly locate lines matching a user-specified pattern called a *regular expression*) can be used with great power. For example, the command `grep '^whereof\ ' *.ct` will print the number of occurrences of *whereof* in all word-count files in a directory, and `grep '^vn.*[iy]nge*\ '` prints all words beginning with *vn-* and ending in some form of *-ing*.

After my experience with the collocation counts and VMS DATATRIEVE, I am convinced that using files of this form and standard UNIX utilities is much more efficient than using a real database system. Eventually I abandoned the database

```
vtter 6 0.045
vtter'd 1 0.008
vtterly 5 0.038
vultures 1 0.008
wade 1 0.008
wafers 1 0.008
wager 4 0.030
wages 2 0.015
waie 1 0.008
waies 23 0.173
waight 9 0.068
wait 22 0.165
waite 7 0.053
waited 2 0.015
waites 2 0.015
waiting 3 0.023
waits 3 0.023
wak'st 1 0.008
waken 2 0.015
```

Figure 2–6: Part of a list of word counts produced using `awk`

system and maintained my counts and results in simple file structures, processing them using UNIX commands and utilities. Database systems are often not very efficient when running on even a moderately-loaded multi-user computer system, unlike utilities like `awk` and `grep`. DATATRIEVE is by no means a horribly bad database system. Still it did not provide the flexibility, power and efficiency I achieved using UNIX tools.

Of course, this approach will not suit someone who does not have access to a UNIX system. In addition, UNIX is notoriously difficult for beginners and non-programmers to use, and the documentation is often poor. As noted at the beginning of this section, the choice of software must be based on one's own experience and the local environment. However, I would strongly advise anyone working with texts and counts from texts to begin with simple data structures in files, using something like the word list shown in Figure 2–6. While obviously simple, the power of an approach based on such lists may not be evident. But

most of the counts and comparisons that form the the heart of this dissertation were produced from nothing more complicated.

2.6 A Quantitative Analysis of the Authors' Use of Contractions

To close this chapter, Shakespeare's and Fletcher's use of compound contractions will be examined. Since word-count lists for both the expanded and unexpanded texts have been created, it is quite straightforward to measure the change in rates due to expansion. This should give some idea of how much effect expansion may have on a study based on word counts. It may also turn up some interesting differences in the two authors' usage, which might perhaps be useful in discrimination. (Note that in the discussion that follows, the use of the word "contractions" will refer to compound contractions, which are formed by joining two or more words.)

First, consider the concept of a ***contraction index,*** a statistic for measuring the change in a rate due to expansion. Several different definitions for such a measure are possible. I have chosen the following one: given two counts, X_O made before expansion ("O" for "original") and X_E afterwards ("E" for "expanded"), the contraction index is defined as:

$$C = 100 \times \frac{(X_E - X_O)}{X_E}$$

The index C is therefore the proportion of the expanded count that was added by the expansion process (expressed as a percentage). X_O and X_E could be the counts for a single word or the total number of words in a text, in which case C would be an indicator of the overall contraction rate.

Word counts for the twenty Shakespeare and the six Fletcher control texts were combined to create overall word lists for each author. The total number of words in these texts was used to calculate an overall contraction index for

each playwright. The two counts and the index C for both writers are listed in Table 2–4, which also includes the individual word counts and contraction indices for the words that are most frequently contracted with another word. As expected, Fletcher's overall rate is larger than Shakespeare's. Inspection of the indices for individual words indicates that he also uses more contractions of each common word. (There are exceptions, including *would* and some forms he rarely uses, such as *wilt* and *art*.) Three words are by far the most commonly contracted forms in both writers: *is, will* and *it*. Fletcher's use of contracted forms of these words is remarkably high; almost two-thirds of the occurrences of *is* are are found in forms like *it's* and *all's*.

While Shakespeare's overall rate of contraction is lower than the younger dramatist's, we would expect that his later plays would show evidence of his adjustment to changing practices among dramatists and his generally more "relaxed" style. Table 2–5 lists the number of tokens (both expanded and unexpanded) and the contraction indices for all 24 Shakespeare and 8 Fletcher plays, listed according to date of composition. *King John* and *Richard II* are the two plays with the least proportion of compound contractions. The overall contraction indices for Shakespeare's plays begin to increase with *The Merry Wives of Windsor* and *Twelfth Night,* which contain large amounts of prose, and reach a maximum of 1.96 with *Coriolanus*. The values for these late plays are in the same range as the values for the eight Fletcher texts.

	Fletcher			Shakespeare		
Word	X_O	X_E	C	X_O	X_E	C
Overall	130879	133649	2.07	421622	426265	1.09
is	793	2037	61.07	5066	6972	27.34
will	660	1406	53.06	2590	3917	33.88
it	1306	2182	40.15	4192	5610	25.28
vs	315	430	26.74	876	1037	15.53
there	394	511	22.90	988	1192	17.11
let	308	414	25.60	1038	1174	11.58
here	291	343	15.16	1113	1221	8.85
where	210	254	17.32	653	693	5.77
he	818	985	16.95	3422	3630	5.73
on	372	452	17.70	1497	1574	4.89
we	601	690	12.90	1775	1959	9.39
was	287	332	13.55	1163	1252	7.11
I	4033	4519	10.75	11084	12033	7.89
has	262	297	11.78	177	188	5.85
who	123	142	13.38	604	628	3.82
in	993	1149	13.58	5557	5763	3.57
were	209	231	9.52	851	921	7.60
she	560	609	8.05	1284	1403	8.48
what	766	829	7.60	2311	2483	6.93
that	1436	1584	9.34	5568	5788	3.80
shalt	20	22	9.09	144	145	0.69
the	2995	3171	5.55	13482	13873	2.82
wilt	25	26	3.85	159	166	4.22
open	35	36	2.78	63	66	4.55
would	443	457	3.06	1229	1283	4.21
hath	30	32	6.25	1035	1036	0.10
of	1658	1722	3.72	8322	8489	1.97
for	1208	1260	4.13	4006	4052	1.14
art	108	108	0.00	420	442	4.98
you	1971	2042	3.48	7561	7672	1.45
to	2583	2615	1.22	9386	9513	1.34

Note: C indicates the contraction index values. X_E is the number of occurrences after expansion, and X_O is the count in the unexpanded text. Entries are sorted by the average C value.

Table 2–4: Contraction indices in Shakespeare and Fletcher control texts

	Word Counts			*is*		*will*		*it*	
Play	X_O	X_E	C	X_E	C	X_E	C	X_E	C
CE	14426	14546	0.82	202	29.70	131	36.64	180	12.22
LLL	20870	21021	0.72	356	15.73	209	22.97	268	13.43
TGV	16926	17114	1.10	317	29.65	165	40.00	244	27.46
R3	28141	28285	0.51	328	18.29	219	27.85	266	20.68
TS	20431	20692	1.26	381	31.50	235	37.02	285	31.58
MND	16158	16219	0.38	208	7.69	142	23.24	143	7.69
Rom	23950	24174	0.93	455	24.40	225	40.00	283	20.85
KJ	20418	20503	0.41	239	16.32	133	25.56	190	14.21
R2	21848	21945	0.44	296	13.18	127	29.13	200	18.50
MV	20958	21098	0.66	316	17.09	181	31.49	265	11.70
1H4	24078	24268	0.78	290	25.17	235	42.98	252	16.27
MAN	20811	20959	0.71	388	21.91	235	20.85	289	11.76
AYL	21306	21465	0.74	355	19.15	224	30.80	251	19.92
H5	24876	25022	0.58	429	18.65	220	19.55	308	16.88
JC	19126	19232	0.55	305	18.03	167	17.96	230	15.22
MWW	21086	21385	1.40	447	27.07	322	35.40	278	27.34
TN	19417	19722	1.55	370	34.86	228	34.65	298	34.56
AWW	22533	22891	1.56	435	35.63	213	32.86	407	31.70
Mac	16070	16336	1.63	300	41.00	112	40.18	238	36.97
Ant	23673	24117	1.84	442	39.14	201	42.29	366	41.80
Cor	26451	26981	1.96	383	43.34	214	43.93	384	39.84
Cym	26683	27188	1.86	464	39.01	220	48.18	426	35.68
WT	24506	24969	1.85	391	38.11	203	47.29	442	39.59
Tem	16025	16302	1.70	252	39.68	152	48.03	196	33.16
Bond	20019	20449	2.10	297	67.00	198	55.56	290	39.66
Priz	22975	23541	2.40	352	65.91	270	64.07	349	38.97
Vale	24454	24844	1.57	353	41.93	170	62.35	348	30.75
Thom	20644	21015	1.77	312	56.41	239	49.79	328	32.62
Chan	16070	16447	2.29	330	58.18	182	51.10	304	46.05
Subj	25493	26011	1.99	369	62.87	287	52.61	453	43.49
Deme	24035	24578	2.21	382	59.95	275	48.73	423	41.61
Prin	22287	22623	1.49	307	52.12	194	43.81	363	30.85

Table 2–5: Contraction indices by play

Table 2–5 also lists the expanded word count W_E and contraction index C for *is, will* and *it*. For all three words, the index values in Shakespeare's late plays fall in Fletcher's range. There seems to be a greater difference in rate of contraction for *is,* however. Fletcher is more prone to use *is* in a contracted form in every situation, as shown by examination of the rates of the compound contractions before expansion. The combined rate for all forms of contracted *is* in the unexpanded texts is 9.24 per thousand for Fletcher and 4.29 for Shakespeare. The majority of these occurrences are *'tis* for both writers: 3.29 for Fletcher, and 1.61 for Shakespeare. The large majority (all but 1.19 in Fletcher, 0.26 in Shakespeare) are enclitic contractions with common words like *that, there, what, all* and personal pronouns.[17]

As noted earlier, Smith prepared standard editions of a number of plays. One of his tables [148, p. 12] contains counts from which the contraction index for *is* can be calculated for 5 Jacobean plays not included in this study:

Play	Author, date	W_O	W_E	C
The Revenger's Tragedy	(Anonymous, c. 1607)	143	278	48.6
Pericles	(Shakespeare?, c. 1608)	143	267	46.4
The Atheist's Tragedy	(Tourneur, c. 1609)	183	378	51.6
The White Devil	(Webster, c. 1612)	245	390	37.2
Women Beware Women	(Middleton, c. 1623)	148	515	71.3

The index value for *Pericles* is larger than in any of the Shakespeare plays in Table 2–5. Middleton appears to be at least as fond of *is* contractions as Fletcher, while Webster resembles Shakespeare in this regard.

Ignoring for the moment the question of possible modifications by scribes and printers, can one use the contraction index for *is* to discriminate between samples

---

[17]It might appear that finding the different forms and rates presented in this paragraph involved a lot of work. In fact, this is an excellent example of the power of the approach that was discussed in the last section. The search utility **grep** was used to search each author's overall word list; words ending in apostrophe-*s* and other contracted forms of *is* were found and put into a file. **awk** was then used to sum the rates in the third column, and the file was modified with an editor to observe and then delete the contractions involving *'tis* and other common words. The entire examination took under three minutes.

Frequency Distribution for the Contraction Index of *is:*

	Shakespeare			Fletcher		
Value of C	Num.	%	Cum. %	Num.	%	Cum. %
0	2	0.9	0.9	0	0.0	0.0
0–6	10	4.5	5.4	0	0.0	0.0
6–12	23	10.4	15.8	0	0.0	0.0
12–18	30	13.5	29.3	1	1.4	1.4
18–24	34	15.3	44.6	0	0.0	1.4
24–30	21	9.5	54.1	2	2.9	4.3
30–36	39	17.6	71.6	2	2.9	7.2
36–42	23	10.4	82.0	4	5.8	13.0
42–48	16	7.2	89.2	8	11.6	24.6
48–54	16	7.2	96.4	9	13.0	37.7
54–60	6	2.7	99.1	7	10.1	47.8
60–66	1	0.5	99.5	16	23.2	71.0
66–72	1	0.5	100.0	10	14.5	85.5
72–78	0	0.0	100.0	7	10.1	95.7
78–84	0	0.0	100.0	2	2.9	98.6
84–90	0	0.0	100.0	1	1.4	100.0
Totals	222			69		
Mean:		27.84			57.87	
Std Dev.:		14.43			14.04	

Note: Contraction index measured in scenes in control and test sets that contain at least 15 occurrences of *is* after expansion.

Table 2–6: Frequency disribution for the contraction index of *is*

of Shakespeare and Fletcher? The word is frequent in both authors' texts (a rate after expansion of 15.2 per thousand in Fletcher, 16.4 in Shakespeare), but from the definition of C one might imagine that one or two contracted forms could greatly affect the index in samples containing a small number of occurrences of *is*. First, C was measured in acts in all 32 plays of known authorship. The values in 109 of the 121 Shakespeare acts (90.1%) are less than 42.0 (about the mid-point between both authors' mean value of C). Only 2 of the 40 Fletcher acts (5.0%) have values this low.

Next, frequency distributions of scenes from the 34 plays of known authorship were examined. It seems wise not to include scenes with a small number of

occurrences of *is*. In the two disputed plays, 25 of the 46 scenes have at least 15 occurrences of *is* (in the expanded versions). Values for scenes in the plays of known authorship that meet this requirement were found, and a frequency distribution prepared. This is presented in Table 2–6; for each range of C values it lists the number of scenes, percentage of the total scenes and a cumulative percentage. Examination of the table shows that over four-fifths of Shakespeare's scenes have values of C lower than 42.0, compared to 13% for Fletcher. There is more variation and overlap of values than when C was measured by acts.

The overlap in the authors' range of values indicates that this measure of the extent of contracted forms of *is* cannot be used by itself as an absolute indicator of authorship. However, statistically this feature is as good a discriminator as the best found in this entire study.[18] While the contraction index for *is* could be used in combination with other variables, no further use of this variable was made in this study. First, the analyses of Chapters 4 and 6 combine variables of the same or similar nature (collocation and proportional pair counts in the first, word rates in the second). Second and more importantly, contracted forms were susceptible to alteration by scribes and compositors, and the index C might be altered rather drastically by one or two changes in a short scene. While no textual feature is immune to possible alteration, the danger in this instance is judged to be larger than for the other features examined in the rest of this study.

Table 2–7 shows the values in the 15 scenes from *The Two Noble Kinsmen* and *Henry VIII* that contain 15 or more occurrences of *is*. In the disputed plays, where some possibility exists of revision of one author's work by the other, judging authorship based on contracted forms seems even more unreliable than in other situations. Therefore, the results in Table 2–7 are interesting but not

---

[18]Obviously this statement anticipates the results of Chapter 4 and Chapter 5. In Chapter 4 the use of t tests will be used to measure the difference between the authors' mean rates in terms of the variance. The values of the statistics for the values of C in the samples that make up the frequency distribution of Table 2–6 are: $t' = 15.41$, $\nu' = 115.7$, prob. $= 2.66 \times 10^{-15}$. This probability is as small as any of those for the final set of markers used in Chapter 5, which are listed in Table 5–12 on page 215.

The Two Noble Kinsmen				*Henry VIII*			
Act/scene	C	X_E	X_O	Act/scene	C	X_E	X_O
I.i	21.05	19	15	I.i	39.29	28	17
I.ii	52.63	19	9	I.ii	29.17	24	17
II.ii	45.00	40	22	I.iv	38.89	18	11
II.iii	81.25	16	3	II.i	31.58	19	13
III.v	55.56	18	8	II.ii	63.16	19	7
III.vi	51.85	27	13	II.iii	47.06	17	9
IV.i	64.29	28	10	II.iv	44.44	18	10
IV.ii	66.67	18	6	III.iia	41.38	29	17
IV.iii	50.00	16	8	III.iib	42.86	28	16
V.ii	69.57	23	7	IV.i	45.45	22	12
V.iii	19.23	26	21	IV.ii	6.25	16	15
V.iv	33.33	15	10	V.i	45.16	31	17
				V.iii	61.11	18	7

Note: C indicates the contraction index value. X_E is the number of occurrences of *is* after expansion, and X_O is the count in the unexpanded text.

Table 2–7: Contraction index for *is* in scenes of the disputed plays

necessarily informative. A number of scenes usually attributed to Fletcher have very high values of C (*TNK* II.iii, IV.i–ii, V.ii and *H8* II.ii, V.iii), and two of the very low values occur in scenes usually assigned to Shakespeare (*TNK* I.i, V.iii). In addition, *H8* IV.ii, usually given to Fletcher, has an extremely low value which is unparalleled in the 69 Fletcher scenes included in the frequency distribution listed in Table 2–6.

Many scenes have values somewhere between the authors' average values. Table 2–5 shows that Shakespeare's use of contracted forms of *is* certainly increases in his late plays, which were written closest to the date of composition of both *TNK* and *H8*. Thus, even if one could trust that contracted forms reflect an author's intentions, this variable would not be as good a marker as these statistics suggest once variation with date of composition was taken into account.

2.6.1 Remarks on the Computing Techniques

The characteristics of old-spelling Jacobean texts certainly provide a challenge for those who wish to use computer software to find and count lexical features. No matter what sort of concordance or word-searching software is employed, the problems of variant spellings and contracted forms will affect results obtained from texts that simply reproduce the copy text. The system of codes and conversion presented in this chapter is independent of the software used to count features in the texts. This procedure for data alteration could be usefully employed to pre-process texts for more sophisticated software (for example, a word-tagging or parsing system).

The procedure has certainly proved successful in this study. The initial development of the coding schemes, the conversion lists and the program REPLACE has been completed. Their existence would certainly benefit any other researcher wishing to use this method with 16th and 17th century texts. Unfortunately the pre-editing stage is unavoidable. Adding hash suffixes to the 34 Shakespeare and Fletcher texts was a long and tedious process, even with powerful search and editing facilities. But the strategy based on coding and replacement probably represents the most efficient way of handling variants and contractions in large samples. In any case, the versions of the texts that result from this procedure play a prominent role in this study.

While the results presented in the final section of this chapter may not have led to any useful findings for the authorship investigator, the study of compound contractions in the two dramatists is interesting in its own right. The analysis of these forms has also provided an illustration of how computing techniques described in this chapter can be used to the search for textual features that discriminate between Shakespeare and Fletcher.

Chapter 3

Some Recent Stylometric Studies

The application of statistical procedures to problems of Shakespearean authorship is not a recent development. The question of Bacon's or Marlowe's authorship of Shakespeare's works was the subject of an early attempts to use a statistical analysis of textual features to resolve questions of disputed authorship. In 1851 de Morgan suggested that average word length could be used to resolve the authenticity questions surrounding the Pauline epistles. An American physicist, Mendenhall, noted de Morgan's suggestion and in 1887 published the results of a study of the frequency distributions of word length in samples of four English writers [84]. He followed this with a comparison of Shakespeare, Bacon and Marlowe; this article, published in 1901, was entitled "A Mechanical Solution of a Literary Problem" [85]. Mendenhall's analysis of these writers' "word spectrums" showed that Shakespeare's curve (unlike that of most other writers studied) had its peak at the four-letter word. His word length distribution was unlike Bacon's in several ways; however, Marlowe's distribution agreed substantially with Shakespeare's.

Mendenhall's pioneer work suffers from the fact that important statistical techniques had not been developed yet (for example, goodness-of-fit tests and the measurement of statistical errors). But he grasped the central concepts of a statistical approach to authorship study. By comparing the variation within samples of an author's work to the differences between writers, textual features

are isolated that can be used to discriminate between writers. This is the basic principle behind all quantitative authorship studies.

The last fifty years have witnessed an increased number of applications of statistical methods in authorship studies. Studies have appeared more frequently as the increasing popularity of computing in literary research has encouraged all kinds of quantitative studies. A scientist approaching an authorship problem can choose between a wide variety of approaches that have been suggested and evaluated by various scholars. These approaches can be fundamentally different regarding both the choice of the textual features to be studied and the statistical procedures to be employed. Holmes' recent article, "The Analysis of Literary Style — A Review" [51], provides an excellent summary of the various statistical approaches to literary analysis.

As noted in the Introduction, this study will focus on function words. It therefore belongs with those studies that attempt to develop authorship methods based on traits that all writers share but use at different and characteristic rates. This chapter will examine several other studies that share this approach. Considerable attention will be placed on research that has used methods related to those developed by Morton, Michaelson and their associates. Their particular approach has been applied to problems involving texts written at different times and in different languages, including several Shakespearean questions. (Chapter 4 will describe an analysis based on these methods using collocations and proportional pairs.) Several other studies that focus on function words or the present authorship question are also discussed, including Mosteller and Wallace's important examination of *The Federalist* papers.

3.1 The Development of Positional Stylometry in English

In the last 20 years, Morton, Michaelson and their associates have proposed classes of textual features as general and reliable indicators of authorship in Greek, Swedish and English texts. If their techniques are effective in the field of Elizabethan and Jacobean drama, then they represent an important new tool in textual studies. Metz, a literary scholar who uses these methods in an analysis of *Titus Andronicus,* goes so far as to state that "Shakespearean authenticity studies may have entered a new era" [94]. The methods have remained controversial, however, and several articles offering conflicting evaluations of their reliability have appeared in *The Shakespeare Newsletter, Computers and the Humanities* and the *ALLC Bulletin.* Reviewing a number of submissions to *The Shakespeare Newsletter,* the editor Marder observes: "Clearly the science or art of stylometry has not been so perfected that the half dozen or so workers in the field of Shakespeare authorship can use it without incurring the wrath of the others" [78].

These techniques of stylometry originate in the work begun in the late 1940s on Classical Greek texts. Wake's initial work on sentence-length distributions led to studies by Morton, Michaelson and others of word mobility and finally "positional" stylometry. In 1974 Morton and Michaelson took these ideas and made the switch to English in an effort to provide evidence for a court of law. An evaluation of the techniques in a number of literary texts led to an examination of Shakespeare's *Pericles* as an illustration of the method's use in testing the homogeneity of a text. This one example of an application led to the examination of several other Shakespearean authorship questions. These results eventually generated similar popular interest and academic controversy to that which followed the results of the analysis of the Pauline epistles a decade earlier.

A brief examination of the development of stylometry in Greek will help one to understand the rationale behind the variables and statistical analysis evaluated in Chapter 4. The development of similar methods of stylometry in English by Morton and his colleagues at Edinburgh will be described. A modification of these techniques and their application to disputed texts in the field of economics by O'Brien and Darnell is examined. Finally, some Shakespearean applications of the stylometric methods based on the original methods of Morton and his colleagues will be reviewed.

3.1.1 Wake's Sentence-length Studies

Morton and his colleagues recognize the publication in 1957 of Wake's "Sentence-Length Distributions of Greek Authors" [166] as the pioneering work in their approach to stylometry. In taking up a study of sentence length, Wake set himself apart from many of his colleagues who turned their attention to vocabulary studies. Sentence length and authorship in English had been earlier examined by Yule and Williams, but the method seemed unsuccessful. Yule concluded that the variability in the known works of an author were significantly large. Afterwards such researchers as Yule and Herdan carried out studies on other aspects of language, and these formed the basis of what one might call the "vocabulary" school of authorship studies. The recent research of Muller and Ule (for example, "Recent Progress in Computer Methods of Authorship Determination" [163]) reflects ideas and methods associated with this approach to the statistical analysis of language.

Wake discovered that Yule's conclusion was based on a miscalculation; the differences between the Coleridge samples he examined were not really significantly different. Wake then began his examination of Greek authors by pursuing Williams' suggestion (first published in 1939 and later described in *Style and Vocabulary* [172]) that sentence-length distributions are log-normal. Although he found that the log-normal distribution fits his data for the most part, he noted several objections to its use. His data contained certain examples that

were clearly more complex in nature (for example, bimodal distributions). In addition several writers exhibited characteristic behavior in the upper tail of the distribution.[1] For these reasons, Wake compared the means and quantiles of his samples (using these statistics' standard errors) in order to evaluate the consistency within a writer and the differences between writers.

Before using sentence length on authorship problems in Plato and Aristotle, Wake tested his method on many samples of unquestioned authorship. The results matched scholarly opinion regarding several texts in the Xenophon and Aristotle canons. A greater degree of variability occurred in some works by Plato. Wake expressed his belief that this result was exceptional because of the nature of the dialogue form, the small size of samples studied, and the observed chronological trend evident in the statistics. In the final section of his paper, Wake compared the statistics calculated from the Plato and Aristotle controls to two disputed works.

A number of characteristics set this study apart from many earlier authorship studies. The first is the extensive use of control samples of known provenance. Wake was also quite clear in describing exactly what he was measuring (sentence length in *continuous* prose), and was aware that changes in the nature of the samples might affect the measurements. Such considerations often played an important role in his sampling methods. He also shows an appreciation that the surviving tenth-century A.D. texts may not reflect the author's original intentions. Sentences were considered corrupt and omitted from the analysis when multiple manuscripts existed and a comparison showed enough disagreement between them to alter the sentence length. Likewise Wake was sensitive to the effects of modern editing on the texts he used in his study. By comparing two

---

[1]More recent studies have also encountered this problem. Morton [108] uses a measure of skewness suggested by Davies to determine when sentence distributions can be fitted to a log-normal distribution. Sichel [141] notes that the log-normal model for sentence lengths should be rejected because the sentence-length distributions observed by Wake and Williams are negatively skew after transformation. He then demonstrates a compound Poisson distribution that fits published distributions very well.

modern editions he demonstrated that the differences so introduced were small compared to the random variation between samples. Wake's attention to these important factors should be studied with care by anyone undertaking an authorship study.

Morton began his work in Greek by building on Wake's findings. Studies of very short and very long sentences led to the examination of word mobility within sentences. He and Michaelson developed techniques based upon the position of common words in sentences, concentrating on the first, second, penultimate and final positions. Positional stylometry was born, and Morton was later to state that "the marriage of frequency of occurrence with position of occurrence has proved to be the key to stylometry" [107, p. 10]. Morton's book *Literary Detection* [108] provides a summary of the methods and results of stylometric studies in Greek with which he has been associated.[2] Many of these studies have elicited strong criticism. Hockey [50] and Oakman [116] summarize the controversies. The Greek studies are of less interest in the current discussion than the manner in which stylometric methods evolved for English.

3.1.2 Habits of Authorship in English

Morton and his colleagues have published a number of works outlining their method of stylometry in English. The first two (both published in 1978) are Morton's book *Literary Detection* [108] and the report "To Couple Is the Custom" by Michaelson, Morton and Hamilton-Smith [102]. The two chapters in Morton's book that deal with collocations in English appear to be a revised version of the joint report. Two more recent papers will also be cited. The

---

[2]Further details can be found in numerous articles and books, including: "The Authorship of Greek Prose" [106], "On Certain Statistical Features of the Pauline Epistles" [70], *Paul, the Man and the Myth: A Study in the Authorship of Greek Prose* [111], "The New Stylometry: A One-word Test of Authorship for Greek Writers" [98], "Last Words" [97], "Positional Stylometry" [99], "Things Aint What They Used to Be" [103], "The Spaces in Between: A Multiple Test of Authorship for Greek Writers" [100] and *The Genesis of John* [110]

first, "The Nature of Stylometry," was prepared by Morton and Michaelson for *Stylometrics '84*, a workshop on authorship studies held at the University of Edinburgh in August 1984. (There are plans to publish it as a technical report of the university's Computer Science Department.) The second paper is the text of Morton's address to the Royal Society of Edinburgh in February 1985, entitled "Fingerprinting the Mind." While these reports may not be familiar to students of authorship studies, they are important because they reflect some new developments in methodology and respond to certain criticisms published since the appearance of *Literary Detection*. In reviewing the development of their method as set out in these works, some of the textual features regarded as indicators of authorship will be described. A discussion of the statistical methods used to analyze the data taken from text samples will follow. Finally, their application of the method to a number of texts and authors in order to validate the method will be examined.

Characteristics of Habits of Authorship

The set of tests for English texts developed by Morton and his colleagues reflect the general principles of stylometry as described in Chapter 1. In several of the publications describing their methods, the authors carefully examine desired characteristics of variables in authorship studies. In *Literary Detection* Morton states that a proposed habit should be "apparent in a choice which frequently confronts all authors." Of course, it also must be something that can be measured and numerically expressed. He recognizes a third attribute for proposed habits, one which is central to using stylometry to solve many authorship questions: "It must be a habit which can be shown to be unaffected by changes in subject matter, by the passing of time, by reasonable differences in literary form and all other possible influences which might affect the habit [108, pp. 96f]."

Frequency of Occurrence

Morton continually emphasizes frequency of occurrence. In his publications he makes several observations regarding the necessity of measuring frequently-occurring features. The first is the desirability of counting features that are spread relatively evenly throughout a text in order to be sensitive to modifications in any part of the text. The goal is really to establish the textual feature's pattern of occurrence, and naturally the more occurrences there are the more accurately one can predict expected values and estimate variation in small samples [112,107]. Also, the analyses of English texts in "To Couple Is the Custom" and *Literary Detection* rely for the most part on χ^2 tests, and most statisticians suggest that this procedure not be used when the number of expected observations in a sample is less than five. The authors of "To Couple Is the Custom" note that this requirement often dictates a minimum length for samples that can be analyzed, but add: "The number of occurrences often needs to be larger than the minimum for another reason, which is that statistical theory applies to events which occur in a random fashion and language is not random in fine detail and can only be treated as random if the samples are large enough" [102, p. 4]. (Discussion of the meaning of the word "random" in this context will be postponed until Section 3.1.4.)

Identifying Occurrences

The second of the three criteria for a habit set forth in *Literary Detection* is that of numerical expressibility. Any statistical method must be based on quantitative measurements that are not dependent on the researcher's subjective or intuitive classifications. But problems of detail arise in counting features in a text, especially when the feature is not clearly defined or when certain characteristics of a given text complicate identification. In stylometric studies these problems usually center on whether an orthographic word-form (in Morton's terms, a word-form) is a variant of a given word *type*. For example, does one count *a'* in a poem by Burns as an occurrence of *all*, or *'tis* as an occurrence of *it* or *is*? These often appear

to be simply questions of definition. But the issues may be complex (as noted in Section 2.2.2 in regard to contractions in 17th century texts), and a decision may seriously affect the final counts.

In discussing the subject of definitions Morton observes that scholars (in this case, of literature or linguistics) often feel uncomfortable with the apparently simplistic and arbitrary definitions that scientists and statisticians impose on language features. The latter group are quite satisfied with usable definitions "supported by a demonstration that the proportion of uncertain, doubtful or ambiguous observations is so small that it cannot affect the judgement based on the observations" [107, p. 10]. Few published stylometry studies convincingly demonstrate that such definitions and assumptions are of no importance to the outcome.

The problem of definition and countability confronts anyone making a textual quantitative study, even if the issue is never discussed in the published results. The easiest solution is to make a rule and stick with it; this is the usual approach to handling homonyms. In their study of *The Federalist Papers* Mosteller and Wallace equate all words of the same spelling (capitalization neglected), even to the extent of lumping the Roman numeral *I* in with counts for the first-person pronoun. They justify this solely on the grounds that it makes routine counting easier (especially by computer), and a more consistent treatment will result [113, p. 16]. Morton observes: "As long as you adopt the brutally simple method of classifying words by their form and not their function, you have reduced the uncertainty of measurement by one degree" [108, p. 31]. But after adopting this convention, Morton breaks it several times in the demonstrations of English stylometry given in *Literary Detection* and "To Couple Is the Custom," usually for sound reasons based on the nature of the text. (Examples are postponed until Section 3.1.4.) The "stick with the hard and fast rule" attitude is often quite unsatisfactory, especially if one fails to anticipate all the peculiarities of a text.

In regard to homonyms, Morton joins Mosteller and Wallace in stating that

the frequency of these situations is so low as not to affect their method. "But the difficulties arise in words which are, by stylometric standards, rare and the proportion of uncertainty is so small that it cannot affect any decision made using the counts of frequent words made by a computer" [107, p. 11]. Section 2.2.1 showed that this is not true in 17th century English texts and even in modern editions of these texts that reflect characteristics of Early Modern English. Pronouns and articles, the most frequently occurring words in these texts, are frequently found in contracted forms and often have unexpected homonyms.[3] The assumptions made and criteria adopted should be made clear in every study and re-evaluated for each application of a method, since they will depend on the characteristics of a particular text. Morton often remarks: "All enquiries start and end with the text" (for example, in *Literary Detection* [108, p. 15]). Unfortunately, this excellent advice has often not been followed by those applying scientific techniques to authorship problems.

Forms of the Variables Used in Positional Stylometry

One of the most innovative aspects of positional stylometry as developed by Morton and his colleagues is the choice of textual features to study. While some of these forms are identical to those used in positional studies in Greek, Morton and his colleagues developed a number of ways of studying position in uninflected English. The trend away from sentence lengths continued, and increased reliance on punctuation-independent position definitions resulted in the large number of collocation tests that have frequently been used in Shakespearean studies. In reviewing Morton and his colleagues' descriptions of their forms of authorship tests, particular examples of these forms are often referred to as "habits" when in fact they may not be "habitual" in the situation being described. In my description of these results I have occasionally adopted this usage; thus "habit,"

---

[3]Indeed, the choices made in such situations have helped bring about differing conclusions in stylometric analyses of the same text. This will be discussed further in Section 3.3.3 on page 121.

"textual feature" and "test" are sometimes used interchangeably to refer, for example, to a particular collocation. The intended usage should usually be clear in context.

The first form of test, a carry-over from the Greek studies of word mobility, measures the use of *frequent words in certain positions in the sentence.* Allowing for the differences between inflected and uninflected languages, Morton anticipated that observation of texts with differing literary forms would show more variability [108, p. 130]. Tests of this form might also be subject to varying degrees of direct and indirect speech [102, p. 16].

While tests based on position in the sentence are less suited to an uninflected language, another form of test makes use of the fact that constituents in such languages are identified by their word order or relationship to words of other classes. *Collocations,* defined here simply as the placing of two words in immediate succession, should be free of the influences of punctuation or sentence length variations. The usefulness of adjacent pairs of words in authorship studies was first noted by Michaelson and Morton when they examined the distributions of word positions measured in blocks delimited by occurrences of common words in a text. Most of the collocations used in positional stylometry studies involve conjunctions, articles, prepositions or pronouns; some are therefore very frequent. In "To Couple Is the Custom" it is asserted that the occurrence of a collocations varies "much less within a writer and much more between writers than the occurrence of either word which makes up the collocation" [102, p. 17].

Collocations are usually measured as the proportion of one of the members of the pair (usually called the *keyword* by Morton and the *node* by linguists) that are followed by (or alternatively, preceded by) the other. Most of the recent work by Morton and his colleagues has only used "followed by" collocations. Smith also favors the latter form because "the natural mode of expression in English is a 'followed by' sequence" [144]. The relationship is usually expressed using FB to indicate "followed by" and PB "preceded by" (for example *and* FB *the*).

Keyword	followed by	Keyword	followed by
it	<adjective>	*in*	*the*
	X *and*	*of*	*a*
	X *of*		*the*
and	<adjective>		X *and*
	the	*the*	X *and*
	X *the*		X *the*
be	*a*		X X *the*
but	*a*	*to*	*be*
by	*the*		*the*
I	*am*		bracketed by verbs
	have		

Figure 3–1: List of collocations from "The Nature of Stylometry"

In this study all collocations are assumed to be "followed by" tests (represented simply as *and the*, for example) unless explicitly noted.

Similar lists of collocations that are often useful indicators of authorship appear in *Literary Detection* and "To Couple Is the Custom." These appear to have been compiled by experience; no mention is made of using the computer to automatically search for collocations that discriminate. "The Nature of Stylometry" contains a somewhat different list, which is reproduced in Figure 3–1. Several examples in this table depart from the simple definition of collocations given above. First, in two cases any "adjective" following the keyword counts as an occurrence of the collocation. Morton quite explicitly provides a "usable" definition of what he means by an adjective [108, p. 137].[4] Second, some collocations are formed by the co-occurrence of two words separated by a specified number of any other words. Such patterns are usually represented by using the symbol "X" to represent an intervening word. For example, "and with the" and "and write the" both count as occurrences of *and* X *the*.

---

[4]This definition is also given in "To Couple Is the Custom" on page 25. The authors apparently assume that classifying verbs will not lead to difficulties since a similar definition for the pattern of "*to* bracketed by verbs" is not provided.

Another form of test, ***proportional pairs*** (sometimes called proportionate pairs), was also adopted from the Greek methods. The idea here is that there are pairs of words in which the number of occurrences of one word is a constant proportion of the number of occurrences of either member of the pair. This phenomenon was first observed by Morton for the Greek adjectives for *all* and *many,* and a number of pairs with similar patterns of occurrence were found in English. Morton suggests (in *Literary Detection,* pp. 146–150, and "To Couple Is the Custom, page 17) that nothing can be inferred about a link of meaning or function and that the pairings are purely observational. The pairings given in *Literary Detection* are obviously related in function and meaning, but Morton's caution is intended to emphasize the fact that not all related words can be assumed to occur in this manner. Only through validation on samples of known authorship can such pairs (and any other putative habit) be evaluated. Examining writers' preferences for one of a pair of words has also been studied by Ellegård in connection with the Junius letters [31,32]. This idea plays some role in Mosteller and Wallace's choice of marker words in *the Federalist* papers study [113]. While proportional pairs are introduced in publications about positional stylometry, they are based on relative frequencies and have nothing to do with word position.

The final form of test used by Morton and Michaelson in recent analyses was developed after the publication of *Literary Detection* and "To Couple Is the Custom." The method measures *positions of once-occurring words* and is the subject of a recent article in the journal *Literary and Linguistic Computing* [109]. Recognizing the difficulties in analyzing the pattern of occurrence of the class of once-occurring words (especially the obvious characteristic that the number of occurrences depends on sample size), Morton simplifies the problems by just measuring where they occur in the text. Position is again measured in two ways, by location in the sentence and in relation to frequent keywords. By comparing the numbers occurring in two positions (for example, first word of a sentence to last word, or the immediate left and right of a keyword), Morton claims to have developed an effective and reliable test.

Clearly computer assistance is necessary here for all but the smallest samples, and objections regarding once-occurring word-forms are likely to arise. How does one treat homonyms, inflectional endings and different forms of verbs? Morton states that an analysis of a number of English and Greek texts indicates that differentiating homonyms only alters the observed counts of once-occurring words by less than half of one percent, and will not affect the results of a study to any great degree [109, pp. 1]. No details of this analysis are provided, and one wonders if this result would hold for old-spelling Jacobean texts with their many variants and homonyms.

3.1.3 χ^2 Tests and the Basic Assumptions

How are the basic stylometric assumptions of consistency with a writer's works and significant variation between writers statistically tested? ~~In the methods of positional stylometry developed in~~ In *Literary Detection* and "To Couple Is the Custom," a number applications of stylometric studies in English are given as examples. In several of these cases the rates of occurrence were so alike or dislike that no further statistical test was performed. In most cases, however, the χ^2 test was used to evaluate the significance of the differences between the measurements.

In testing the hypothesis of consistency, samples for a given writer were divided into a number of smaller samples. Assuming that these smaller samples of a writer were from a larger population (that is, all of the writer's works), the overall counts for the set of samples were used to estimate the mean for this population and then the expected number of occurrences in each sample. Finally, the χ^2 test was performed to determine whether the differences between the observed and expected values were significant (usually at the 5% level). Non-significant χ^2 values were taken to indicate consistency.[5]

---

[5]By statistical convention, uppercase Greek letters are usually used to represent

In a similar way the χ^2 test was used to show that differences between writers were significant. Problems involving the assignment of an unknown or disputed sample to an author were solved as follows in "To Couple Is the Custom." A set of habits that were consistent in samples of the author was found. Counts for these control samples were then pooled and compared to the counts in the disputed sample using a series of χ^2 tests, each with one degree of freedom. If the problem involved choosing between two candidates for authorship, the same procedure was repeated for the second author. χ^2 values that were significant at the 5% level were interpreted as support for a difference in authorship, and the question was resolved if the proportion of significant values from the series of χ^2 tests was either high or low enough that the decision could be made by inspection. In each of the four English literary examples given in *Literary Detection* and "To Couple Is the Custom," the interpretation of the series of χ^2 values was obvious. But what procedures are specified when an assignment by inspection of these values is not in order?

By using the χ^2 test in this manner, one is testing the Null Hypothesis that ~~the~~ each sample has identical proportions of keywords classified and not classified as occurrences of the given habit. Testing a disputed sample against the counts for one given author, the null hypothesis can be reformulated thus: are the counts independent of the classification by author? Morton and his colleagues look for significant χ^2 values in order to assert that the differences observed are large enough that the samples must come from two populations with different rates of occurrence. If analysis shows that rates in samples by a single author do not significantly differ, they conclude that the disputed sample must not be by the author of the control sample.

A Type I error in this example would be that of accepting the result implied

---

statistics calculated from data, and lowercase letters are used for the corresponding distributions. In this study, however, I will follow Snedecor and Cochran [151] and Bailey [4] in using the symbol χ^2 to represent both the family of distributions and the statistic.

by significant χ^2 value when there was no difference in authorship; the probability of such an error is equal to the significance level chosen (usually 5%). By using a number of tests and producing a series of χ^2 values, Morton and his colleagues hope to minimize this likelihood. In "To Couple Is the Custom" they imply that the probabilities resulting from each χ^2 value can be multiplied together to yield an overall probability of error, if the tests are independent [102, pp. 12 and 67]. These products of probabilities have been used in a number of studies by Morton and others to calculate likelihood ratios. An overall probability of error is calculated from a series of tests for each candidate for authorship, and the ratio of these values is interpreted as the likelihood that Author A wrote the sample rather than Author B. Examples of this method of probability combination include studies by Morton of *Titus Andronicus* (described on page 116) and Merriam's work on the Huntingdon plays (described on page 126).

Even assuming that there is little or no correlation between tests, the statisticians I have consulted in this university have been very reluctant to accept this use of likelihood ratios without an underlying multivariate model to describe the pattern of occurrence of these features.

The statistical independence of tests was evaluated by Morton in twenty chapters of Scott's novel *The Antiquary.* (The total size of these samples seems to have been around 100,000 words.) He calculated Pearson product-moment correlation coefficients for a number of combinations of tests, starting with tests of a similar pattern (that is, involving the same keywords or sentence position):

> Proceeding in this way to check for correlation between tests of the same pattern and then checking the different patterns against each other, it is clear that no statistically significant correlation exists to any greater degree than chance expectation. It is therefore justifiable to take these tests as being independent of each other [108, p. 142].

The reference to "chance expectation" refers to an earlier statement that of the 351 total combinations of tests "17 pairs would be likely to show correlation significant at the 5% level and 3 correlation at the 1% level." It is unclear why he introduces the use of levels of significance (which are normally associated with

using a significance test on a given hypothesis) at this stage, and one can conclude from his statement that some pairs of tests did show significant correlation. In any case, these results seem to be the only published data concerning the correlation of individual tests in English. Every subsequent study that assumes that such tests are independent appears to rely on this single observation of their pattern of occurrence in one nineteenth century novel.

The combination of information from a number of tests has been a subject of much debate in recent stylometric studies. Merriam has summed χ^2 values from a series of 2×2 tables, treating the result as distributed according to χ^2 with the degrees of freedom equal to the number of tests summed [89]. This additive property of χ^2 is well-known, although Smith has recommended developing multidimensional contingency tables [149]. The manner in which O'Brien and Darnell [117] have combined information for collocations based around a single keyword is similar to Smith's proposal. Brainerd notes that tables with a higher number of degrees of freedom result in a more powerful test [19], but he mentions this in the context of grouping cells in a $m \times n$ contingency table. In his detailed discussion of various χ^2 tests he does not demonstrate how to combine a series of individual, unrelated tests.

Recently Morton has suggested privately that a better way of combining the results of a number of χ^2 tests is that outlined by Fisher [36] and Kendall [60]. Given n independent tests of significance, each yielding a probability p_i, the statistic

$$-2 \sum_{i=1}^{n} \log_e p_i$$

is distributed as χ^2 with $2n$ degrees of freedom. Merriam makes use of this statistic in a recent examination of data originally presented by Smith, although Smith appears unimpressed with its validity [91]. Once again this calculation assumes independence and the examples given in Kendall's book involve the repetition of the same test on different data, unlike the current problem of combining data from different tests on a single sample. In private discussions Morton and Michaelson have recognized these problems and agree that it is not clear how

to combine probabilities resulting from a series of χ^2 tests validly when the statistical relationship between these textual features is not fully understood.

3.1.4 Anomalies

In giving examples of stylometric tests and their pattern of occurrence in a number of English texts, the authors of "To Couple Is the Custom" discuss a number of complications that sometimes arise when testing samples. It is not claimed that each test described will work for every writer; indeed, the lists provided are simply frequent patterns that are often effective. For any given author the tests will fall into three categories: those that are consistent within that writer's works; those that vary significantly and are therefore useless; and finally, those that "were found to be consistent within most works of a writer but showed occasional anomalies which would prevent them being used as a test of authorship until the reason for the anomalies had been brought to light" [102, p. 21].

This final class is further divided into those tests in which the inconsistencies occur in a "periodic" manner and those in which the increased variance is due to a peculiarity of the text or a portion of the text. The first type of anomaly can often be accounted for by modifying the statistical procedure, but the second requires that the researcher recognize that authors can at times introduce perfectly reasonable phrases or repetitions that alter the general pattern of occurrence in a text. This implies that one should always investigate the readings in a sample that actually give rise to the counts. In large studies this makes a researcher responsible for examining what may be thousands of counts, a burden that many would be unwilling to accept. Avoiding this responsibility is even easier now that computers can take over much of the tedious job of searching and counting, which makes possible projects of immense scale that would otherwise be impractical. To their credit Morton and his colleagues continue to emphasize that one should carefully examine the occurrences of textual features that give rise to statistical results.

Anomalies due to characteristics of the text are not rare. Often direct or reported speech affects the number and location of personal pronouns. Morton and his colleagues have noted occasions when the two proportional pairs *no*+*not* and *this*+*that* are inconsistent within a writer's works unless separated according to their grammatical function [102, p. 23] (another blow to the idea that identifying words only by form will not cause problems). Common phrases that border on formulas often characterize a writer or a character. Such phrases, such as ("used to be" or "a sort of"), can significantly affect the counts of a particular collocation. Elements repeated in a list can cause local concentrations of such collocations as *the* X *and* or *the* X X *the*. A character in *The Antiquary* is constantly referred to as "The Old Man;" this enhances the counts for the collocation of *the* followed by an adjective. Many of these types of anomaly occur when a pattern or usage that is normally functional becomes tied to style or content. The exceptions to the general rule are not seen by Morton to disprove the rule but to remind one that the tests may not always reflect an author's subconscious pattern of usage.

Perhaps for many the term *periodic* conjures up images of sine waves drawn on a school blackboard. In *Literary Detection* Morton uses the term to refer to the idea of *serial correlation,* where the occurrence of an event changes the probability of a subsequent occurrence [108, p. 84]. Statistically known as *contagion,* this effect often leads to a pattern of occurrence where events come in runs or small clusters separated by lengthy gaps. For example, it is often observed that writers repeat constructions (such as a particular sentence beginning) for stylistic purposes [102, p. 22].

Three methods of dealing with contagion are suggested. First, runs (sequences of consecutive occurrences of a feature) can be counted as a single occurrence. Effectively the textual feature being counted is redefined. The second method is to increase the minimum length of the samples being used until the effects of runs and clusters disappear. "Usually samples which are twice the minimum size, i.e. which contain ten occurrences rather than five occurrences,

will be free from periodic fluctuations" [102, p. 22]. The third way of dealing with contagion is to use a distribution such as the Poisson, hypergeometric or negative binomial to calculate expected values. Of the three methods, the final one is probably the most desirable because it attempts to give the most complete description of the data.

3.1.5 Habits and their Statistical Distributions

This discussion of periodic leads to an important assumption about the tests under consideration. Obviously rates of occurrence will vary in samples belonging to the same population, and the extent of this variation depends on the distributional model underlying the occurrence of the language features. In developing procedures to measure "consistency" within an author's works, one must make assumptions about the limits of variation that will be accepted. Morton and his colleagues assume that most of their tests are characterized by binomial distributions, in which occurrences of events are independent of each other. "It has been so far assumed that these occurrences fit the simplest pattern, the binomial, a pattern appropriate to a mutually exclusive choice for which the rate of occurrence is unchanged from trial to trial" [102, p. 31].

The description of stylometry in "To Couple Is the Custom" indicates that when a textual feature is not binomial in an author's texts, the basic methods described may indicate significant differences between samples when they are actually by the same hand. The test is then labeled "anomalous" and is not considered to be a useful habit of authorship for that writer unless allowance can be made for the more complex pattern of occurrence. There is no underlying theory that asserts that all habits must be binomial. In "The Nature of Stylometry" (in the section "Refinements and the Complications") Morton states:

> These habits are biological material. There can be no question of arguing that any writer, or speaker, must produce some statistical

> pattern for any habit. Normally composers will conform to particular patterns, but being human beings, they can always create exceptions.

The importance of the binomial model is not stressed well enough in "To Couple Is the Custom." The authors do not verify that the majority of the tests they employ are distributed binomially and only introduce the topic of distributions to explain significant χ^2 values for certain anomalous habits in Scott. The distributional assumptions underlying the method have not been recognized by others adopting these techniques, judging from the fact that no other stylometric study has attached any great importance to the binomial distribution. Yet Morton and his colleagues make it clear that the binomial model is closely associated with the use of χ^2 tests for validating the method's basic assumptions in control samples.

How can one make use of distributional models other than the binomial? Recall that for features which are characterized by the binomial distribution, consistency within an author's works was evaluated using the χ^2 test on the individual samples. The examination in "To Couple Is the Custom" showed that some habits in Scott's novels exhibited significant χ^2 values when counted by chapters. These habits were then counted in equal-sized blocks of keywords, and the results fitted to the Poisson distribution. If a χ^2 test of goodness-of-fit indicated that the data was indeed Poisson then this was taken to imply consistency within authors [102, pp. 29–31]. Non-binomial habits were not discovered in any of the applications in which a disputed sample was to be assigned or rejected, and thus Morton and his colleagues did not demonstrate how non-binomial habits could be used in discriminating between authors. The only comment seems to argue that this will not be necessary:

> In any actual case the procedure will be determined by the number and extent of the samples available, the differences between the texts which are to be examined, and the scale of the project. All that need be shown here is that there is a plenitude of material on which to work. If the samples are large and the study of them is meant to be detailed and definitive, then all the habits would be recorded and

> their sampling distribution investigated and the distribution appropriate to each would be established and employed. But if the texts were to be no more extensive than the samples set out in this table [of habits in *The Antiquary* and *Castle Dangerous*], and if the investigation was secondary to some more important process, then a small number of habits might well prove decisive especially if some combinations were compiled, such as asking what proportion of sentences have as their first word either *a* or *and* or *the* or *but*. There is no point in shooting a corpse and if simple tests are decisive it makes no sense to go further, unless to improve the techniques of testing [102, p. 39].

Perhaps it is unfortunate that the authors did not go further "to improve the techniques of testing" in the case of features that are not distributed according to the binomial pattern. Merriam [92] shows that a number of collocations in the Shakespeare First Folio are not distributed according to the binomial distribution, and results presented in Chapter 4 show that the basic χ^2 testing described above does not produce satisfactory results in comparing Shakespeare and Fletcher. The uncertainties regarding the combination of individual tests and the statistical distributions of individual habits are the most serious questions facing those who wish to apply these authorship methods. Examination of a more complete sample of an author's works should help to determine whether these problems seriously affect this application of stylometric principles.

3.1.6 The Extent of Testing and Validation

As described in *Literary Detection* and "To Couple Is the Custom," the proposed habits and the statistical methods were applied to a number of literary texts in an effort to validate the method for general use in English. The premise that these habits were consistent within a writer's works was tested in the following samples, three of which were chosen because there were stylistic or external reasons to expect internal variations. (Pages 132–136 of *Literary Detection* provide a discussion of the choices.)

1. Ten prose samples of different twentieth-century English authors, each roughly a thousand words.

2. Two samples from *The Antiquary* (1816) and *Castle Dangerous* (1831), each over 50,000 words. About 15 additional chapters seem to have been used from *The Antiquary* for some purposes [108, pp. 135f].

3. Two chapters each from two novels by Henry James: Chapters 1 and 2 from *The Americans* (1877); and Chapters 1 and 2 from *The Ambassadors* (1903).

4. Samples from three novels by John Fowles written between 1963–66: Chapters 1, 2, and 60 from *The French Lieutenant's Woman*; Chapters 1, 2, 77, and 78 from *The Magus*; and thirty-five pages from *The Collector*.

Token counts for the samples by James and Fowles are not provided, but rough estimates based on the relative frequencies of some common words are 25,000 words for the James samples and 30,000 words for Fowles. The total number of tokens in the texts initially used in validating the hypothesis of internal consistency is then approximately 215,000 words.

Having established to their satisfaction that the tests did produce the expected results in these samples, Morton and his colleagues turned to several problems of authorship to show that the method could be successfully used to resolve such questions. A brief description of each problem and the conclusion follows:

1. Some Shakespearean scholars have proposed that Acts 3–5 of the play *Pericles* are authentic Shakespeare but that Acts 1–2 are by another hand. In comparing the habit counts in the two sections, the authors found "no statistically significant difference in any preferred position or collocation" [102, pp. 62–65]. However, the collocation of *to* followed by a verb is marked with a note in the table stating: "The distribution is Poisson and the difference is not statistically significant." (Applying the χ^2 test to the counts given confirm that the probability associated with χ^2 is significant.) Another collocation, *to the* has a significant value of 5.63 for

1 degree of freedom, but no additional comment is made. (Smith has used similar techniques to reach a different conclusion about *Pericles*; this will be discussed in Section 3.3.3.)

2. Jane Austen's unfinished novel *Sanditon* was anonymously completed by an admirer referred to as the "Other Lady" in 1975. Two chapters from *Sense and Sensibility* and two from *Emma* were compared with two chapters from Austen's share of *Sanditon* to establish internal consistency in her works. Comparison with two chapters of the novel by the Other Lady revealed nine habits with significant differences. The authors of "To Couple Is the Custom" suggest that this result strongly supports the claim that these markers of authorship are subconscious and cannot easily be imitated, even by someone who succeeds in reproducing another person's style.

3. Another test of the success of imitators was the comparison of samples from two novels written as "new" adventures of Sherlock Holmes by a pair of modern writers. Provided with information (from the two writers) that certain chapters of these collaborations were the sole work of one imitator, Morton demonstrated that stylometric techniques distinguished both men's samples from a genuine Holmes story. In addition, when provided with an article written by one of the imitators he showed that one imitator's habits were significantly different from this control, while the other's were not. Thus he correctly matched the article with the first writer's chapters [108, pp. 192–194].

4. Finally, Morton and his colleagues tested the proposition that six unattributed articles in the nineteenth century *Fraser's Magazine for Town and Country* were written by Elizabeth Gaskell. These six were compared to two attributed works, and no significant statistical differences were revealed. The conclusion was that there was "no obstacle to the hypothesis that these papers are by Mrs. Gaskell" [102, pp. 77–86].

In the studies in which the internal consistency of an author's works was tested, a number of anomalies (as described in Section 3.1.4) were discovered and taken into account. A description of these provides further insight into the occasions when a habit might fail to behave as expected and thus reflects to some degree the robustness of the method on actual data. In testing for homogeneity within Scott's *The Antiquary*, five habits with significant χ^2 values are found. When three of these were counted in equal-sized blocks, the observations fit the Poisson distribution and are thus taken to be consistent. The collocation *of the* exhibits the characteristics of serial correlation, and when Morton counted sequences of occurrences instead of individual occurrences the χ^2 test showed no significant differences between observed and expected values. The anomaly for the occurrence of *and* as the first word of a sentence was traced to varying amounts of direct speech in the chapters; when the chapters were taken in pairs the significant differences vanished. A minimum sample size criterion of ten expected occurrences in a sample (rather than the five usually recommended for χ^2 tests) was used for tests in Scott. Therefore, these results might suggest that local variations due to periodic effects will occur even in samples with ten expected occurrences.

Two problems with the habits in *Castle Dangerous* are relevant to the application of these methods to dramatic texts since they concern the use of dialect. For the two proportional pairs *any+all* and *no+not* the occurrence of the Scottish contracted forms *a'* for *all* and *no'* for *not* were responsible for a significant statistical difference. This vanished when the counts for the variants were pooled with the standard forms. Once again the convention of recognizing words only by their form suffers at the hands of an inconsiderate author.

Another way in which a textual characteristic affects the counting of a habit is revealed in the analysis of the collocation *it is* in Fowles' novels. Interestingly enough the explanation of this anomaly in *Literary Detection* differs from that in "To Couple Is the Custom." Morton's book indicates that the difference is due to different proportions of past tense usage within the texts. Counting *it*

followed by *is* or *was* results in a habit that is consistent with the three novels [108, p. 144]. On the other hand, the explanation in the technical report indicates that the abbreviation *it's* is very common in *The Magus,* and that if the counts for this contraction are added to those for *it is* the differences are no longer significant [102, pp. 54f].

Scholars are allowed to change their minds, and there is no reason why both modifications of the counts might not yield a consistent result. Both explanations for this anomaly emphasize the point Morton makes in *Literary Detection* about *it is* in Fowles: "It must be kept in mind that the aim is to produce tests of authorship and for these statistically significant differences in habits to be explained only by a difference in authorship, not a change in genre, or of literary form, or of historical perspective." But the adoption of a "brutally simple method of classifying words by their form" gives rise to difficulties even in the applications by which Morton and his colleagues attempt to validate their methods. These examples give further support to the idea that these textual features should be regarded as lexical rather than graphical characteristics. They should also remind a researcher that definitions of occurrence and non-occurrence may indeed affect results. Detailed examination of the effects of variant forms and contractions would seem to be a justifiable precaution in most studies.

3.2 O'Brien and Darnell's Modified Method

In 1982 O'Brien and Darnell published *Authorship Puzzles in the History of Economics: A Statistical Approach* [117] which contains a study of six authorship problems in economic literature. These two researchers based their study on a modified version of the stylometric methods outlined in the preceding section. Additional details of their method (and response to some criticisms of the book) appears in a paper entitled "A Statistical Technique for the Investigation of Authorship Puzzles" [118][6]

O'Brien and Darnell's studies are of interest for a number of reasons. Their research represents an application of the general techniques developed by Morton and his colleagues to new and different problems. While their method relies on tests of collocations and the position of frequent words in sentences, they do not calculate a χ^2 value for each test but combine related tests, thereby avoiding the difficulty of interpreting a large number of probabilities. Their method proved less successful for a 17th century problem than for 19th century texts because of differences in the use of high frequency words. Finally, they used a Monte Carlo computer simulation to evaluate the power and error rate of their method; in addition, the simulation was used to study the minimum sample size for which the method performed acceptably.

The introductory chapters of O'Brien and Darnell's book provide an extremely lucid description of the nature of authorship problems and statistical approaches for solving them. The authors emphasize their belief that statistical analysis of internal evidence should be carried out in conjunction with an independent evaluation of external evidence to best resolve a question of authorship. In doing so they criticize Bayesian techniques in which prior odds (based on the

---

[6] I am grateful to Prof. O'Brien for providing me with a copy of this report after I sent him a letter querying some aspects of their study.

researcher's interpretation of the external evidence) are necessary for the analysis of internal evidence. The most important example of such an approach is Mosteller and Wallace's study of *The Federalist Papers*. O'Brien and Darnell's criticism of this study is unfair; their description of Mosteller and Wallace's discussion of prior odds is incomplete and misleading. (See Section 3.4.1 on page 129 for further discussion of Mosteller and Wallace's methods and this criticism.) O'Brien and Darnell's other objection to Bayesian methods assumes that "it is necessary to specify a particular distribution of the characteristic in question" [117, p. 16]. Certainly the main study in Mosteller and Wallace's book [113] uses distributional information to determine the posterior odds of authorship, but this is supplemented with a robust[7] Bayesian analysis that makes no use of distributional information. (This method performed less satisfactorily than the main study, however.) Techniques that make use of distributional models do not have to be Bayesian, and there is no reason why a Bayesian approach could not be used in a distribution-free method. But in addition to these arguments supporting their use of non-parametric methods, O'Brien and Darnell state (and their work attempts to demonstrate) that simpler techniques are sufficient.

3.2.1 χ^2 Tests and Classes of Habits

While retaining the theory developed by Morton and Michaelson, O'Brien and Darnell substantially alter the methodology. Instead of calculating a series of χ^2 tests (essentially from 2×2 tables), they combine related tests and reduce each comparison of samples to only three χ^2 values calculated from $n \times 2$ contingency tables. A row in such a table now corresponds to an individual test in the earlier procedure; a column corresponds to each sample being tested.

Only three such tables are used. The first is composed of the counts of common words beginning a sentence. The other two tables are based on collocations

---

[7] In statistics a *robust* procedure is one that is not very sensitive to departures from distributional assumptions (usually of normality).

O'Brien and Darnell's method: $n \times 2$ Contingency table for first word of a sentence evaluated with a χ^2 test

	Text A	Text B
Number of sentences beginning with Word 1	x_1	y_1
Number of sentences beginning with Word 2	x_2	y_2
etc.	*etc.*	
Number of sentences beginning with Word n	x_n	y_n

Morton and colleagues' method: The sequence of 2×2 tables corresponding to the $n \times 2$ table

	Text A	Text B
Number of sentences beginning with Word 1	x_1	y_1
Number of sentences NOT beginning with Word 1	$n_x - x_1$	$n_y - y_1$

Number of sentences beginning with Word 2	x_2	y_2
Number of sentences NOT beginning with Word 2	$n_x - x_2$	$n_y - y_2$
etc.	*etc.*	

	Text A	Text B
Number of sentences beginning with Word n	x_n	y_n
Number of sentences NOT beginning with Word n	$n_x - x_n$	$n_y - y_n$

Figure 3–2: Use of contingency tables: O'Brien and Darnell vs."To Couple Is the Custom"

involving the keywords *be* and *the*. These are among the most common words in English and are often preceded by other frequent prepositions, conjunctions and pronouns. By counting the "preceded by" collocations (that is, the number of occurrences of *the* preceded by *of* rather than the number of occurrences of *of* followed by *the*), a $n \times 2$ contingency table is produced yielding a single χ^2 value with $n-1$ degrees of freedom. Such a table is illustrated in Figure 3–2; the corresponding sequence of 2×2 tables that is used in "To Couple Is the Custom" and *Literary Detection* is also shown. Note that in Morton's method the number of keywords marked by a habit (such as "first word of sentence") is always compared to the number that are not marked. In O'Brien and Darnell's method, this second value is not always included in the contingency table. If they have made use of this total, then it is explicitly listed as a row in the table. (Examination of the tables listed in *Authorship Puzzles in the History of Economics* shows that this "all other" category was used about half the time.)

3.2.2 Amalgamation of Counts

Another major departure from the method described in "To Couple Is the Custom" concerns O'Brien and Darnell's amalgamation of counts. They aggregate some rows of data to meet the statistical requirements of the χ^2 test (that is, 80% of the cells of a contingency table should have expected values greater than five) and to group habits that are consistent within authors but do not discriminate [118, p. 8]. In *Authorship Puzzles in the History of Economics* they plainly state that they "have tried, at least as far as possible, to amalgamate cells in a way which makes literary sense." As an example of "literary sense" at work they list *may be* and *might be* as habits that could be grouped but assert that *must be* should not be grouped with either. This is an unfortunate example because in one of their applications they actually do amalgamate counts for *must* and *may* in one group, while another group contains the count for *might* [117, pp. 10 and 20]. Indeed, one must wonder if the section in the introductory chapters explaining the rationale behind amalgamation was written before the actual

analyses, or perhaps the confusion could be explained by the dual authorship of the book. The statement on page 10 that "as far as is possible, we have avoided such amalgamations" appears unjustified when the actual tables are examined. Many tables are composed of rows with counts for as many as six habits aggregated without any comment on the choice of groupings. In the example mentioned earlier for collocations with *be,* a table with four rows is produced:

1. *can*
2. *would, to*
3. *might, could, will,* any noun
4. *must, should, may, shall, need*

This grouping scheme is not atypical. The literary or linguistic sense which dictated these choices is not obvious and is not explained. Nor is their definition of a noun outlined; a footnote seems to indicate that pronouns are to be included in this category [117, p. 212].

Perhaps responding to such criticisms as these, the authors explain their rationale more clearly in the working report [118]. They state that they did not amalgamate to achieve maximum discrimination but that the groupings used in their book were the first that were observed to be used consistently within authors and differently between authors. Thus the sets of groupings seem to have resulted from intuition, eye-balling of the counts and trial and error. But O'Brien and Darnell stress that the same amalgamations are tested for internal consistency for each candidate and discrimination between the candidates. Only combinations of habits and groupings that satisfy both these two criteria are used to assign disputed texts. O'Brien and Darnell describe a strict procedure for choosing both the set of habits and their groupings for a given problem. While the use of this objective testing procedure is admirable, one questions whether they examine enough texts of known authorship to establish an author's characteristics. Internal consistency within an author is evaluated by comparing only two samples (often halves of a single work). This definition of internal

consistency is probably too simplistic, and the results of their study would be easier to accept if they had examined more text samples by each candidate.

O'Brien and Darnell see a number of advantages of their methodology compared to that described in "To Couple Is the Custom." First, constructing a table for a given class of habits makes some allowance for statistical dependence between the individual habits in that class. Also, the tests then use "all the available information on a particular class of habit ... simultaneously rather than sequentially," resulting in contingency tables with more than one degree of freedom. The focal point of the testing is shifted from individual habits to a broad class of habits (such as first words of sentences or collocations of *the*). Each of the three class tests produce a single statistic, simplifying comparisons and reducing the problems associated with the combination of results from a number of tests [117, p. 29]. The authors have anticipated questions of statistical independence. "But, in the last resort, the proof of the pudding is in the eating. If the method works — and it seems that it does — then over-scrupulous objections seem rather otiose" [117, p. 11]. This attitude is also evident in much of the work of Morton and his colleagues. The responsibility of demonstrating the validity of the method *in the texts being studied* lands squarely on the shoulders of the researcher.

Problems with Some 17th Century Texts

O'Brien and Darnell's methods worked least well when applied to the seventeenth century writings of Child. First, they found that the punctuation of the texts forced them to abandon sentence position tests. Some of Child's writings contain entire paragraphs of almost 150 words that are broken only with commas and semi-colons (which a modern editor would often alter to a full stop). Even if one accepted these as legitimate sentences, the number in a sample of a fixed number of words would usually prove too small for statistical analysis. Perhaps more surprising was the discovery that the 17th century texts they examined contain relatively fewer occurrences of the common words in English [117, p. 41].

In the attribution study involving Child, this severely affected their test based around collocations of *be;* the keyword did not occur often enough in the texts they were using to allow testing for internal consistency.

O'Brien and Darnell give 59 per thousand words as a typical figure for *the* in 17th century texts and 86 as typical in 19th century texts. The corresponding figures for *be* are 11 and 18. For comparison to the texts used in the current study, the figures for *the* and *be* in twenty plays of Shakespeare are 32.5 and 8.4 per thousand; in six plays by Fletcher the rates are 23.6 and 8.5 per thousand. Mosteller and Wallace list rates for a number of common words in samples of Hamilton, Madison, Jay, James Joyce and the King James Bible [113, Section 8.1]. The lowest rate for *the* is 57 in Joyce's *Ulysses*. The King James Bible, published in the early 17th century, has a similar rate (84.7) to Hamilton, Madison and O'Brien and Darnell's figure for 19th century texts. Jay's rate (67.5) is noticeably lower than that of his two contemporaries (91.3 and 93.7). Although the rate of *be* in the Bible (8.9) is slightly lower than typical figure given by O'Brien and Darnell, the rate in *Ulysses* (3.3) is much lower than in any of the samples. Thus the low rates observed by O'Brien and Darnell for these words may not necessarily be a characteristic of Early Modern English. The rates may simply vary in texts according to authorship and genre no matter when the texts were written.

3.2.3 The Monte Carlo Simulation

In an attempt to justify their procedure, O'Brien, Darnell and Peters [118] used a computer program to create a large number of counts corresponding to the three habit-classes used in their method. For example, to evaluate the test of first word usage they specified 14 different population means (corresponding to the 14 words commonly found beginning sentences) for two "authors," A and B. A set of samples by the "unknown" author C are created using the same parameters as A for each habit frequency. Using a multinomial distribution to generate counts for 500 samples representing these distributions, they used their

procedure of χ^2 tests to evaluate internal consistency within all three writers, discrimination between A and B and the ability of the method to recognize that texts by C should be attributed to A. In the simulation for each habit class they found that the error frequency was within the bounds determined by the nature of significance testing. The frequency of wrong assignments was very low.

To study the effects of amalgamation of habits, the tests were also carried out for contingency tables in which rows with infrequent values were combined. Rows were repeatedly combined and each contingency table evaluated until the original table was reduced to only three rows. This process was then repeated for another sequence of row combinations. Although the power of the test did diminish somewhat as more rows were amalgamated, the rate for successful classification of the "unknown" samples remained high. To gain insight on the minimum sample size for each class of habits, the researchers repeated the entire series of tests for 500 samples of progressively smaller size. In this manner they determined that 100 sentences were required for the first-words test, 150 occurrences of *the* required for testing collocations of *the* and 50 occurrences of *be* required for testing collocations of that keyword.

O'Brien, Darnell and Peters felt reassured enough by these results to state that "the Monte Carlo results represent a clear validation of our approach" [118, p. 16]. However, the value of the simulation depends on how accurately the computer-generated counts model occurrences in a writer's texts. Unfortunately this question is never examined, and there are good reasons for believing that their model may not be accurate. The pseudo-random number generator they used produces values that are independent of previous values. The simulation does not reflect the often-observed effect of contagion in the occurrence of language features. Therefore the minimum sample sizes determined by the simulation might not be large enough if the habits show significant contagion. To take account of contagion or correlation in a simulation would require the development of a model for the occurrence and interaction of the individual habits that make up a contingency table in O'Brien and Darnell's method. As described

above they reject distributional techniques in favor of distribution-free methods (in this case χ^2 tests). In using a simulation to evaluate this method they *have* made use of a distributional model, namely the multinomial, and they offer no conjectures about how well this model might fit large samples of a writer's work. Their attempt to validate their method is seriously flawed by their failure to demonstrate the correspondence between their computer simulation and actual observed occurrences in literary texts.

3.3 Positional Stylometry and Shakespearean Studies

The positional stylometry methods developed by Morton and his colleagues have received a significant amount of attention so far in this study. This is warranted on several grounds. First, a set of general techniques to authorship problems, validated on texts of several genres and periods, is an attractive concept. Second, several scholars have applied these techniques to Shakespearean authorship questions since the examination of *Pericles* by Morton and his colleagues. These studies have not convinced everyone, and the controversy has generated criticism of the textual features studied and the statistical methods employed. This section will describe these applications to Elizabethan and Jacobean dramatic questions and the concerns raised by some critics.

3.3.1 Metz and Morton: *Titus Andronicus*

Another play in the Shakespeare canon was soon the subject of the techniques used for *Pericles*. *Titus Andronicus* was examined after Metz, a member of the team preparing the New Variorum edition of the play, contacted Michaelson and Morton at Edinburgh. The results of this study are described by Metz in an article in *Text: Transactions of the Society for Textual Scholarship* [94] and more briefly in the *Shakespeare Newsletter* [96]. In this joint research Metz appears

to have advised the Edinburgh team on textual matters, with Morton reporting the results of the statistical analysis to the article's author by letter.

The problem of *Titus* is similar to that of Pericles; some scholars have suggested that the first act is by another writer, possibly Peele. Seventeen sentence position, collocation and proportional pair tests were used to compare Act I to Acts II–V, and Metz reports that "no statistically significant difference in habits occurred between the two parts." The accompanying table, however, shows that the χ^2 value for one test, *and* followed by an adjective, is significant at 4.27 with 1 degree of freedom. Accepting the homogeneity of *Titus,* the authors compared counts for the entire play to those for *Pericles.* For the 21 tests Metz reports that "there is no significant difference in any habit occurring frequently enough to allow determinative tests to be made." Again this proves not quite accurate when the accompanying table is examined. The counts of *the* as first words of a sentence are 10 of 644 (1.6%) in one play and 41 of 560 (7.3%) in the other. The associated χ^2 value is extremely high at 24.57. It would have been interesting if this anomaly had been examined and explained by Metz or Morton.[8]

Satisfied that the two plays are internally consistent and by a single hand, the next step was to more positively identify that writer. Metz selected *Julius Caesar* as a "touchstone" for Shakespeare; all of *The Arraignment of Paris* and three acts of *David and Bethsabe* were chosen to represent Peele's works. *Julius Caesar* was selected because the source text is excellent, its date of composition falls midway between the early *Titus* and the late *Pericles* and because the contrast between its direct style and the more rich and elaborate verse of the other two plays might challenge the stylometric method.

The identity of the author of *Pericles* and *Titus* with Shakespeare was demonstrated through a series of 35 χ^2 tests on the 3×2 contingency table corresponding to the habit counts in the three plays. Three were significant, the highest

---

[8]I checked both of the significant results described in this paragraph, and the results presented in the paper are accurate.

was *it is* with a value of 13.02 for 2 degrees of freedom. (The others were *is the* and *the* followed by an adjective.) Accepting these three significant results as anomalies, the counts for the three plays were then combined and compared to the total counts in the two Peele texts. The number of significant χ^2 values that resulted were interpreted as indications that these two sets of texts were written by different playwrights. The results of six tests with values significant at the 1% level are listed, and Morton is quoted: "the probability that the works of Peele belong to the same population as the three plays of Shakespeare is less than one in ten thousand million." No explanation of this calculation is given, but the figure appears to be the product of the probabilities associated with the significant χ^2 values.

The study as described by Metz in the *Text* article is somewhat unsatisfying for a number of reasons. It does not adequately explain some of the details of the accompanying tables. The series of tests and conclusions, each leading on to another test, misses out certain comparisons that might prove interesting. A sceptical reader would like to see individual acts of *Julius Caesar* compared to the two Peele texts or individual Peele acts tested against authentic Shakespeare. Showing that the combined counts for three plays by one author differ from combined counts for seven acts by another is not a convincing demonstration that the method is sensitive enough to be able to always classify a single act correctly.

Another source of unease is the introduction of new collocation, proportional pair and sentence position tests at every stage of the argument. If each of these were subjected to the rigors of validation described in *Literary Detection* and "To Couple Is the Custom," the article does not mention it. For example, three of the six tests that distinguish Shakespeare and Peele (*as*+*at*, *me* preceded by *to*, and *of* followed by *this*) are first introduced in analyzing the consistency of *Julius Caesar*, *Pericles* and *Titus*.

Metz concludes that stylometry needs to prove itself many more times before being accepted as a general tool. However, regarding *Titus* he asserts that

"future commentators will be hard pressed to deny the play to Shakespeare." One of his final comments describes the establishment of the the playwright's stylometric habits based on data taken from the entire canon as a "fundamental desideratum for Shakespeare studies."

3.3.2 Merriam, *Henry VIII* and *Sir Thomas More*

Past research conducted by Merriam has addressed the same authorship questions as this study. In three articles published *The Bard* [89,90,88] Merriam describes an analysis of *Henry VIII* based upon collocations, proportional pairs and frequent words in preferred positions. He did not test each scene of the play individually, but began by combining the Shakespearean and non-Shakespearean scenes (as assigned by Spedding) and comparing these two samples. Calculating χ^2 from 2×2 contingency tables for twenty tests, he summed these values, obtaining an accumulated χ^2 value with 20 degrees of freedom. As this value was not significant at the 5% level, Merriam repeatedly re-grouped the individual scenes and performed the calculations again until he maximized the χ^2 value between the groups. Using this new division of the play, he compared his proposed Shakespearean portion of the play to samples from six other Shakespeare plays using the same χ^2 method and another test based on the binomial distribution.

In the second part of the article Merriam discusses what one might deduce about Shakespeare's political and religious views given this new assignment of scenes. He also asserts that the non-Shakespearean scenes were written by three hands, Fletcher, Massinger and another unidentified writer. These authors' hands were identified using small samples from their plays in conjunction with the method used to identify the Shakespearean scenes. Merriam's division is given in Table 3–1. The theory that Massinger was responsible for parts of *Henry VIII* was endorsed by Boyle in the 1880s and by Sykes in the early decades of this century; both of their analyses of internal evidence have been rejected by scholars.

Prologue		unassigned
Act 1	Scene 1–2	Shakespeare
	Scene 3–4	Fletcher
Act 2	Scene 1	Shakespeare
	Scene 2–3	Fletcher
	Scene 4	Shakespeare
Act 3	Scene 1	Shakespeare
	Scene 2 (ll. 1–203)	Shakespeare
	Scene 2 (ll. 204–372)	Fletcher
	Scene 2 (ll. 372–459)	Massinger
Act 4	Scene 1	Massinger
	Scene 2	Shakespeare
Act 5	Scene 1 (ll. 1–55)	unassigned
	Scene 1 (ll. 56–176)	unknown
	Scene 2	unknown
	Scene 3 (ll. 1–113)	unknown
	Scene 3 (ll. 114–181)	Fletcher
	Scene 4–5	Fletcher
Epilogue		Fletcher

Table 3–1: Merriam's division of *Henry VIII*

Merriam next published a study of *Sir Thomas More* which claimed that 90% of the text was Shakespeare's [86]. The analysis and attribution of the several hands in the manuscript has been extensively studied. In 1923 a number of distinguished scholars published a volume of reports providing palaeographic, bibliographical and critical evidence that the additions by Hand D are Shakespeare's autograph.[9] Merriam compared counts for 12 proportional pairs in *More* to counts from Spevack's concordance; in addition, 20 collocations were also counted in the manuscript and in a number of Shakespeare samples (all of *Julius Caesar, Titus Andronicus, Hamlet* and *King Lear,* plus random samples from 26 other plays). He summed all the χ^2 values calculated from the 2×2

[9]Literary scholars seem to be unanimous in their praise of this study. An article by Bald describing the study has been been reprinted in *Evidence for Authorship* [33]; Schoenbaum [125] provides a useful summary.

tables corresponding to each test to produce a χ^2 value with 32 degrees of freedom. The resulting probability was well above the 5% level, and Merriam bases his claim for a common authorship on this evidence.

This represents the acceptance of the Null Hypothesis that Shakespeare wrote the play. The probability that this conclusion is incorrect (a Type II error) cannot be determined by the use of significance tests. Smith maintains that this is reason enough to conclude that Merriam has failed to produce enough evidence to successfully refute scholarly opinion [91]. But to support the validity of the procedure, Merriam compares the Shakespeare counts to Munday's *John a Kent and John a Cumber* and to the anonymous *Edward III*. (Two-thirds of the *More* manuscript is in Munday's hand.) The probabilities calculated by Merriam indicate that the counts for both plays are very unlike the data from his Shakespeare control. He makes no attempt to reconcile his conclusion regarding *More* to the textual characteristics of the manuscript. While Merriam's results certainly contradict the assessments of a number important traditional studies, Metz describes various views claiming a greater role for Shakespeare in the play's composition [95].

3.3.3 Smith's Evaluations and Criticisms

The most detailed analyses of Morton's and Merriam's Shakespearean studies have been provided by Smith. In an early study published in *The Bard,* Smith investigated Hoffman's theory that Marlowe wrote all of *Hero and Leander,* which might provide some support for his responsibility for Shakespeare's texts [150]. He also employed Morton's techniques in an analysis of "A Lover's Complaint" [143]. A second *Bard* article examines Morton's study of *Pericles* [144].

In the latter study he compared the two parts of *Pericles* and found that a significant value of χ^2 resulted when the 20 tests used in Merriam's *Henry VIII* studies were combined. Morton used these same 20 tests in addition to another 10 collocations in one of his published analyses of the play [102]. Smith notes that the significant difference in the combined χ^2 value is due to only two tests,

and as first word of a sentence and the collocation *to the.* Morton's and Smith's results for the first test differ because Smith includes colons and semi-colons as sentence terminators. A significant result for the latter test is present in the tables in "To Couple Is the Custom" but is not discussed; it is not one of the tests used on the play in *Literary Detection.* Smith also notes that, if contracted forms of *to the* are included, the proportions in the two parts are even less alike. Smith introduces further tests not found in earlier studies which further support his conclusion there are significant differences between the two parts of *Pericles.*[10]

Morton's and Smith's different conclusions about *Pericles* have fueled a heated debate on stylometric techniques based around these features. *The Shakespeare Newsletter* appears to have offered its pages as a forum for discussion by Smith, Merriam, Morton and Marder, the editor of the journal.[11] These articles contain varying mixtures of scholarly analysis, differences in definitions and strained tempers, which make for interesting reading and certainly should have boosted subscriptions. The articles published by Smith in the *ALLC Bulletin* [149] and in *Computers and the Humanities* [148] on the whole provide better descriptions of his criticisms of previous studies.

In the first article Smith describes studies of word and sentence length in *Pericles, Hero and Leander, The Rape of Lucrece* and *Venus and Adonis.* He

---

[10]In a later study Smith does not count collocations that span sentence breaks [149]. Morton does not take punctuation into account when counting collocations. This may account for other differences in their counts.

[11]Marder describes Morton's early studies in "Stylometric Analysis and the Pericles Problem" [74] and "Stylometrics: The New Authorship Weapon" [75]; his analysis of the controversies are "The New Disintegration or Reintegration of the Shakespeare Canon" [72], "Scholars Dispute Pericles Data" [73], "Stylometry: Possibilities and Problems" [76] and "Stylometry: The Controversy Continues" [78]. Metz's contributions concern *Titus Andronicus* [96] and *Sir Thomas More* [95]. Merriam's work on *More* is described by Marder [77] and detailed by Merriam himself in "Did Shakespeare Write *Sir Thomas More*?" [87]. Smith has published detailed criticism of Morton's Shakespearean analyses in the *SNL* [147,145,146]. Morton [140] and Merriam [93] have responded in turn to Smith's criticisms published in the *SNL* and in *The Bard.* Recently another round of the discussion was published in the journal *Literary and Linguistic Computing* [91].

concludes that Mendenhall's method based on word length appears "to be so unreliable that any serious student of authorship should discard it." He states that sentence length measures work somewhat better but do not provide sufficient evidence to be used alone in a study.[12] Smith concludes the article with a discussion of χ^2 methods of Morton and Merriam, questioning the reliance on 2×2 tables. He suggests the use of multi-dimensional contingency tables for tests that are based on the same keyword (such as *to be, to the* and *to a*). He reports his observations that proportional pairs can vary considerably in samples of an author's work, even when they are composed of "the least context-dependent words."

Smith proposes that probabilities calculated from the combination of χ^2 tests should not be interpreted literally [149, p. 80]. He notes that features of composition are context-dependent to some degree and that effects of style, characterization or dialogue may alter an author's "habit:"

> While his basic habits appear to remain, some modification arising from these and other influences can be expected. The effect is often unrecognizable and therefore incalculable and may induce variation greater than that associated with authentic random sampling. An interpretation of the values of chi-square as probabilities can therefore be misleading.

---

[12] In this article Smith continues with an assumption he makes in his earlier *ALLC* article [149, p. 77] that problems associated with sentence punctuation in 17th century texts can be avoided if modern editions by a single editor are used: "Given the text, an editor, if revising punctuation, will superimpose his own characteristics" [144, p. 172]. While Wake carefully examined the practice of editors of Greek prose [166], Smith provides no evidence to substantiate this claim. The extent to which copy text punctuation is modified or preserved will depend on the editorial guidelines adopted by the individual editor or the general editor of a series.

Dr. Norman Sanders has kindly provided me with photocopies of the guidelines that he has received from three general editors during his editorial career. Comparison of these guidelines for multi-volume series of Shakespeare's plays (the Revels Plays, the New Penguin Shakespeare and the New Cambridge Shakespeare) shows that principles do vary slightly (for example, regarding colons in the copy text). Using modern editions by one editor is not a practical solution, and the choice between using modern and old-spelling editions has been discussed at the beginning of Chapter 2.

This last sentence is probably an accidental mis-statement or a printing error; one presumes that he is suggesting that the probabilities associated with χ^2 values should not be literally interpreted as propositions of degree of belief. Instead he proposes using the χ^2 statistic as a comparative measure to be interpreted qualitatively. To test whether two parts of a text (such as *Pericles*) are by the same writer, he suggests comparing the χ^2 results between these parts with the statistics for two "new" samples, formed by "combining portions of the original samples so that both contain the same proportion of words by each suspected author." If the researcher judges that this second χ^2 value is much less, then the presence of two hands in the text is supported.

Smith thus views χ^2 tests in a literary context as rough indicators of similarity, which must be subjectively interpreted. He states that the results of these tests should be evaluated in light of one's "experience with the behavior of such tests when applied to the particular literary genre under study." The principle of verifying variables and testing procedures on control samples underlies this last proposal but he does not incorporate it into the testing procedure in any objective or formal manner. As Merriam points out [92, pp. 277–278] a procedure based on scientific and statistical methods cannot conclude with a subjective analysis of the resulting statistics. In any case such a procedure assumes a particular sort of problem: a text where the possible division between authors is known in advance.

Smith's article "An Investigation of Morton's Method to Distinguish Elizabethan Playwrights" [148] is certainly misleadingly entitled. The tests developed by Morton were intended to represent a general method for such studies in English, and the *Pericles* problem was chosen as one of several examples. Smith's investigation concentrates on the method employed by Merriam to divide *Henry VIII* and to attribute *Sir Thomas More* to Shakespeare. This paper raises many serious questions about the variables and procedures used in studies based on Morton's approach to stylometry.

Smith pays due attention to one important feature of 17th century dramas

neglected by earlier studies: contracted forms. For each text studied, he prepared a computer version with all contracted forms expanded in addition to a version exactly reproducing the modern edition he chose as a source. Since expanded forms sometimes contributed significantly to the total counts of words and collocations, he based his results upon the versions of the text files in which contracted forms had been expanded to their full forms.

Merriam's earlier work is closely scrutinized by Smith in this article. Considering *Sir Thomas More,* Smith uses Merriam's counts for the Shakespeare control and tests it against a reduced version of text in which those additions that have been identified (as Heywood's, Dekker's and Chettle's) have been removed. (This reduces the size of the sample from about 20,000 words to 18,560.) If Merriam's theories are correct, this shorter sample should be more purely Shakespearean. Using Merriam's tests and method of combining probabilities, a significant χ^2 value results: this would indicate that Shakespeare was not responsible for the bulk of the play.

To test Merriam's procedure in dividing *Henry VIII,* Smith divided Shakespeare's *The Winter's Tale* into 14 parts. Again using Merriam's variables and procedure, he tests all possible combinations of 7 parts against the remaining 7 parts (1716 total combinations). About one in every three combinations produced a χ^2 value that was significant at the 5% level. Noting that large values were not exceptional occurrences in this sequence of tests, Smith also shows that the largest χ^2 value corresponding to given combination produced a value of 60.92 (for 18 degrees of freedom), roughly comparable to Merriam's value of 66.35 (for 20 degrees of freedom) for his division of *Henry VIII*. He concludes that the internal variation of these features within *Henry VIII* is no greater than that found in *The Winter's Tale*. Unless one proposes multiple authorship for the latter play, Merriam's evidence for collaboration is unconvincing.

In his study of *Sir Thomas More,* Merriam used a different set of variables, choosing 31 collocation and proportional pair tests. These do not rely on punctuation and may therefore be more suitable for dramatic verse and old-spelling

texts. Adopting this set of tests, Smith compared five other Jacobean plays (*Pericles, The White Devil, The Atheist's Tragedy, The Revenger's Tragedy* and *Women Beware Women*) to the many combinations of the parts of *The Winter's Tale*. Smith uses the power of the computer to evaluate literally thousands of 2×2 contingency tables for each habit; however, his summary of the results is not altogether clear. He shows that no single test consistently and successfully distinguishes the Shakespeare samples from the other texts, but when all tests are combined "there is perhaps a tenuous affinity between the two plays of which Shakespeare wrote all or part" [148, p. 9]. Smith also examines the individual collocations and proportional pairs in detail, raising a number of sensible-sounding objections to a number of them. For example, the proportional pair *do* and *did* might reflect a stylistic change of tense rather than a difference in authorship.

In this study Smith painstakingly reproduced the methods used by Merriam, attempting to discredit both the textual features counted and the statistical method, which he clearly attributes to Morton. One wonders if any other statistical study in any field has evaluated so many 2×2 contingency tables (with and without Yates' correction, I might add). Certainly grave doubts are raised about the general pattern of occurrence of collocations and proportional pairs in these dramas. These center around the question of internal variation. How large a difference in two sets of counts can we expect to find when small parts of plays are combined and recombined? Four plays used in this study have not been used in computerized stylometric research before, and one hopes that someone will make constructive use of these texts to study the internal variation of textual features within plays in comparison to the differences between the Jacobean dramatists.

3.3.4 Merriam, Information Theory and the Huntingdon Plays

Merriam's most extensive stylometric study is "The Consonance of Literary Elements with Mathematical Models: A Study of Authorship in the Huntingdon

Plays" [92]. This dissertation includes an examination of the various objections to Morton's stylometry, and traces a number of misunderstandings that have contributed towards unfair criticism of his work. The meat of the study is the application of tests of collocations and proportional pairs to the two "Huntingdon" plays; external evidence points towards a collaboration by Munday and Chettle. To justify the method, Merriam applies similar tests to samples of Shakespeare and Fletcher, and uses the results to assign authorship for scenes in *The Two Noble Kinsmen.* As will be demonstrated in Section 4.1.2, these efforts are not convincing, mainly because Merriam does not allow for differences in the use of contracted forms by the two writers when counting collocations.

Merriam asks, "Are collocations random variables?" He presents frequency distributions from the Shakespeare First Folio for 11 collocations (counted in blocks of 10 keywords) and shows that the observations fit the negative binomial (and in one case, the Poisson) distribution [92, p. 116–130]. With these results he answers "yes" but makes no use of the distributional model in the analysis that follows, falling back on the significance tests based on 2×2 contingency tables. Indeed, the original claim for collocations in "To Couple Is the Custom" is qualified with the restriction that the features were distributed according to the binomial distribution, with its smaller variance [108, p. 31]. Merriam's results suggest that Morton's assumption that the majority of his "habits" are binomially distributed may not be justified.

Merriam's attributions of sections of the Huntingdon plays relies on the counts of habits in small samples. To avoid statistical objections to the use of χ^2 tests with small numbers, Merriam convincingly argues for the use of Fisher's exact test over the use of χ^2 with or without Yates' correction. However, as noted on page 89, Morton's criteria for a minimum sample size are not simply founded on statistical grounds.

In attacking some of the ideas presented by Smith [149], Merriam states that "the establishment of sample size by rigorous means is essential" [92, p. 279]. His analyses and results would be more convincing if he had determined such a

minimum size by applying his methods on samples of various sizes taken from Elizabethan plays of known authorship. Instead he incorporates ideas of *information content* and *redundancy* from the field of information theory into his method in order to compensate for the effects of "noise" and local anomalies in small samples. He assumes that Shannon's calculations of the redundancy in English "may be taken as a rough guide to the requirements in language in its widest sense" [92, p. 196]. He does not support this assumption with experiments on control samples of Elizabethan drama using the tests he employs in his study of the Huntingdon plays. While these concepts are attractive, their value in the study of literary features cannot be evaluated without extensive validation in samples of known provenance.

3.3.5 Remarks

Stylometry as developed by Morton and his colleagues has produced several forms of authorship test that have been used and evaluated by other researchers. The encouraging results of O'Brien and Darnell in economic problems contrast with Smith's negative findings in disputed Jacobean samples. In the earliest studies of sentence position tests, collocations and proportional pairs, a considerable effort was made to demonstrate that procedures based on these variables produced the correct results in works of known authorship. The results of these validation experiments, described in *Literary Detection* and "To Couple Is the Custom," were based on prose samples by a number of writers of the 19th and 20th centuries, totaling under 250,000 words. Although these stylometric techniques have since been applied to problems in 17th century texts, no corresponding validation of variables or methods in Jacobean texts has been undertaken.

The questions raised by Smith and others relate to the basic assumptions of stylometry: habits exist which vary between authors to a greater degree than they vary within a single writer's samples. Collocations, proportional pairs and sentence position tests were shown to have great promise in reflecting these

characteristics by Morton and his colleagues. However, the questions posed by the nature of Jacobean drama and the small size of disputed samples in *Henry VIII* and *The Two Noble Kinsmen* are not examined in *Literary Detection* and "To Couple Is the Custom." To determine if these techniques can be usefully employed in a study of these two plays, an extensive examination of texts by Shakespeare and Fletcher must be carried out.

3.4 Other Studies

Before examining the effectiveness of some of the techniques of positional stylometry, several other studies of authorship will be reviewed. All but one of these studies focuses on word-rate variables; the exception examines words labeled according to grammatical class and function. For the most part each of these studies belong to the category of authorship studies that are based on traits that all writers share but use at different and characteristic rates. The first study to be reviewed is one of the most important statistical authorship studies published, Mosteller and Wallace's examination of *The Federalist* papers.

3.4.1 Mosteller and Wallace and *The Federalist*

The most complete description of Mosteller and Wallace's authorship study is their 1964 book *Inference and Disputed Authorship: The Federalist* [113]. The two statisticians' main goal was to compare two different approaches to statistical discrimination: the classical method, as developed by Fisher, and the less popular methods of Bayesian inference. They viewed the question of Hamilton's or Madison's responsibility for the 12 disputed papers as a practical vehicle for a case study of discrimination techniques. In this problem, a large number of known texts by Hamilton (94,000 words) and Madison (114,000 words) were used to determine odds of authorship for each disputed paper. Their results show that Madison is very likely to be the author of all twelve papers.

The variables used in the study are simple word frequencies. The initial set of 165 words to be considered was chosen from three sources. First, a list of 363 function words (compiled by Miller, Newman and Friedman) was examined. After eliminating some words regarded as being contextual (such as most pronouns and cardinal numbers), Mosteller and Wallace included the 70 most frequent function words in their pool of words. They also included a random selection of 20 additional words. None of these were selected because of their discriminating ability. (This is important in their statistical analysis, since they use these words to estimate general characteristics of word occurrences.) The second source of potential markers was a screening study of low-frequency words, which yielded 28 variables. Thirty-two papers (divided into three groups) were examined, and words found in one author's papers that rarely occurred in the other's were isolated. This procedure recognized words like *while* and *enough,* favored by Hamilton, and *whilst,* favored by Madison. The third source of potential markers resulted from an examination of a computer-generated list of word counts. Binomial probability paper was used, and 103 words with at least a three standard-deviation difference in the authors' average rate were added to the pool.

From this initial set of 165 words, a final set of 30 words was chosen for use in classifying the disputed papers. Words in the initial set with little discriminating potential were eliminated on the basis of an *importance measure* calculated for each word from the authors' average rate (and based on distributional assumptions, discussed below). Finally, some words deviated significantly between known works by Madison, and these were also discarded. The final set of 30 included 9 high-frequency function words (*also, an, by, of, on, there, this, to* and *upon*).

Throughout these selection procedures Mosteller and Wallace eliminated words "by the hundreds" (p. 39) which they thought might be affected by context. They admit that these decisions were often made on an *ad hoc* and intuitive

basis (p. 18). In the end they discarded all forms of personal pronouns and auxiliary verbs because of fears of contextuality. However, many of the words in their initial set of 165 potential set of markers seem unsafe: *danger, expense(s), city* and *destruction,* for example. Some of the words in the final set of 30 markers fall into this category (*apt, language, probability*). However, the final analysis of the known and disputed works shows that the function words turned out to be the strongest group of markers in the final set. The best individual marker, *upon,* is nearly as strong as the 21 non-function words put together. These results are perhaps the strongest published evidence that "filler" words in language can be powerful discriminators, more valuable than striking personal preferences that involve less-common markers (such as the *while/whilst* distinction for Hamilton and Madison).

While a number of studies have made use of Mosteller and Wallace's results and techniques regarding choice of variables, the statistical methods they employed have not been widely reproduced. The methods are very complex, and (as noted above) *Inference and Disputed Authorship* is more a detailed investigation of statistical techniques than a step-by-step description of how to solve an authorship problem. Their "main study" is a Bayesian discriminant analysis, based on using the negative binomial distribution to describe the occurrence of words in texts. For comparison and validation they present three other methods: a classical linear discriminant analysis, a robust Bayesian analysis and a simplified rate study on "somewhat classical lines." These last three methods are often applied to a simplified version of the problem (for example, only texts of the same length are used). Although they support the general findings of the main study, each suffers from technical weaknesses (p. 264). This discussion will focus on the main study, since it is clearly the most important. (In addition, some aspects of this approach characterize the procedures developed in Chapter 6.)

Bayes' theorem provides a mechanism for combining evidence from prior information with observed data; the data is used to transform a *prior* probability

to a *posterior* probability. Mosteller and Wallace use Bayes' theorem in two different ways. The first use is directly concerned with the question of authorship. The probability that one author wrote a disputed paper is calculated from a critic's subjective analysis of historical or stylistic evidence and from the evidence of word rates in a disputed paper. The latter information is determined using a distributional *probability density function* to calculate the likelihood that a paper of a given length by a particular author contains the observed number of occurrences of a marker word. Mosteller and Wallace showed that word frequencies in the texts are well-described by the negative binomial distribution, which fits the observed distributions more accurately than the simpler Poisson model. They calculate odds of authorship for each individual word, and they multiply the resulting probabilities together, assuming independence. Later they use estimates of the effects of correlation (and other factors) to adjust their results.

The most common argument used against Bayesian methods is that the prior probabilities cannot often be determined easily and objectively. As noted earlier, O'Brien and Darnell criticize Mosteller and Wallace's approach for this reason, maintaining that internal evidence should be evaluated separately from external evidence. However, Mosteller and Wallace recognize and discuss this problem in their study. They develop their methods so that the prior probabilities are only introduced once the internal evidence is evaluated. Their goal is to determine how much one should change one's beliefs about the authorship of a paper, and their hope that they can "produce such strong statistical evidence as to overwhelm any moderate assessment of initial odds" (p. 50) is fulfilled.

Mosteller and Wallace's second use of Bayes' theorem concerns the estimation of parameters for the probability density functions; they state that this second use is "the core of our development" (p. 58). Parameter estimation is an important area of statistics. Non-statisticians are often only familiar with the usual method of calculating a sample statistic from large amounts of data and then using this as a point-estimate for a population parameter. Mosteller and Wallace's second use of Bayesian methods centers on establishing a frequency

distribution for the parameters of the word-frequency distributions. They call the set of parameters for this second distribution *underlying constants* to distinguish them from the parameters of the word-frequency distribution. Using this method to estimate the parameters of the word-frequency distributions allows one to determine how the results could be affected by inaccuracies in parameter estimation. Mosteller and Wallace also make use of negative binomial and Poisson distributions in these estimation methods.

Much of the statistical difficulty of *Inference and Disputed Authorship* stems from the complicated nature of this Bayesian method of parameter estimation. Other more simple methods of estimation could be used with the negative binomial model for word frequencies to determine posterior probabilities of authorship,[13] but it appears that no one has done this. Another problem is the difficulty involved in using the negative binomial distribution, which requires two parameters. Mosteller and Wallace note that the common methods for estimating one of these parameters require that all text samples be of equal length. Even once the parameters have been estimated, the calculation of the likelihood ratio of authorship for a single word is extremely complicated (Sections 4.1 and 4.4 of their book).

In summary, Mosteller and Wallace's study should be studied carefully for the general techniques and principles they present. The basic approach behind their statistical analysis is clearly explained in the first few chapters of *Inference and Disputed Authorship,* but the details of the calculations that follow are most likely too complex for anyone who is not a professional statistician. Their results indicate that common function words are potentially more valuable discriminators than less-frequent markers, a result that has not been appreciated by some literary scholars reviewing this study (for example, Oakman [116] and Hockey [50]).

---

[13]Thompson provides an introduction to various methods of estimation in Part III (Vol. 2, No. 2) of his *ALLC Bulletin* series on literary statistics [161].

3.4.2 Other Studies Based on Word-rates

Ellegård's studies of the Junius letters [29,30] appear to have inspired Austin to use word-rate variables in an Elizabethan authorship question [3]. Greene was reported to have written *The Groats-worth of Wit Bought with a Million of Repentance* (which contains the famous reference to Shakespeare as "an upstart Crow") shortly before he died in 1592. At the time, several other writers were suspected of having a hand in this libelous document. Austin attempts to demonstrate that the true author is Chettle, who saw the manuscript through the press.

Austin based his study on about 100,000 words of Greene's works and 40,000 of Chettle's. (*The Groats-worth* contains about 11,000 words.) The texts were carefully (and manually) examined and edited before computer analysis. Compounds in open form were joined or hyphenated (for example, *mean while*); examples of *tmesis* ("how greatly soever") were reunited; some spelling variants were modernized (*then/than, lest/least, etc.*). During the counting procedures, Austin combined data for singular and plural nouns, inflected verbs and comparative adjectives. In some cases, if the two writers' use of a word differed in all forms, these were brought together in a "root-group." In other cases, where a difference existed in respect to single sense or part of speech, the individual form was retained as a marker. Like Mosteller and Wallace, Austin did not examine pronouns and many verb forms because of worries about contextuality.

Austin used the *distinctiveness ratios* to find marker words favored by each author.[14] He establishes certain criteria for internal variation, and eventually produces 29 markers for Chettle (such as *aim, bewray* and *brook*) and 21 for Green (such as *admire, assure* and *beseech*). No sophisticated statistical procedures are used to compare the use of these words. The rates for each set of

---

[14]Distinctiveness ratios were introduced by Ellegård as a measure of a difference in rate of use between groups of text. This measure us discussed in more detail in Chapter 5, where it will be used in an examination of the vocabularies of Shakespeare and Fletcher.

markers are pooled, and the combined rates for the Greene and Chettle sets in *The Groats-worth* are presented alongside the rates in the controls. Similar analyses are presented for the use of a few characteristic words (such as *ye* and *however*), word-order inversion, various prefixes and suffixes and other linguistic features. In each case, the rates in *The Groats-worth* are closer to those for the Chettle control.

Austin also looked at some high-frequency function words, commenting that one would not expect such "linguistic small change" to discriminate very well (p. 25). Counts were made in blocks of 1000 words in the control and disputed samples. Again, statistical tests were not used to analyze these occurrences. The counts in each author's control set were listed in a frequency distribution, and the values for blocks from *The Groats-worth* compared to each distribution and the relationship interpreted.

Austin's care in the preparation of the texts and his analysis of a number of features is admirable. However, the fact that he does not use objective statistical methods to interpret his results is certainly a fault. The internal variation of his low-frequency markers probably requires a more rigorous analysis. (The combined rate in Chettle for his set of 21 markers is 8.44 per thousand; Greene's rate for his set of 29 markers is 9.42). Waldo, reviewing this study in *Computers and the Humanities* [167], points out that all Austin's evidence may be suspect because of the choice of texts. The "modern" editions that Austin used date from around 1880, and he should be relying on the most authoritative early editions. In addition, Austin has only used a fraction of Greene's works. (He used all the Chettle works that have survived.) Waldo examined several other Greene works, finding examples of many of the traits that Austin claims are Chettle characteristics. Thus, for reasons unrelated to Austin's use of statistics, the value of the procedures he develops and the resulting authorship claim are difficult to evaluate.

A more statistically rigorous analysis of word-rates was used by McColly and Weier [80] in an effort to determine if the so-called *Pearl*-poet was responsible for the five

Middle English poems often attributed to him. Again, their study pays considerable attention to the words to be counted. Using computer parsing routines, they distinguished word forms according to their form classes and meanings. Since no control texts of known authorship exist in this problem, McColly and Weier chose "all presumably content-free words" with a rate of at least one per thousand in at least one of the five poems.

McColly and Weier's statistical analysis calculates a likelihood ratio that two works are by the same author. This statistic is based on the assumption that the word-frequencies are distributed according to the Poisson distribution and assumes independence between words. The method indicates that the five Middle English works are not by the same author. However, application of the procedure to counts from *The Federalist* papers indicates a difference of authorship where none exists for several pairs of comparisons. But the authors conclude that the method does show promise and can be used to indicate common authorship.

However, the assumption of Poisson distributions, upon which their method is based, is extremely suspect. In the statistical appendix to the paper, the authors attempt to justify this assumption. Although they note that studies by Mosteller and Wallace and others indicate that the Poisson does not fit some observed distributions, they assert that the basic "axioms of Poisson variates are reasonably met" and conclude that the Poisson is a valid approximation. Another factor that dictates this choice is that their method of calculating a likelihood ratio cannot be used with any distribution requiring more than one parameter (like the negative binomial).

However reasonable the assumptions for a Poisson distribution might seem, the matter could have been easily resolved if McColly and Weier had carried out goodness-of-fit tests for their observed counts in the five Middle English works. This is not difficult, and one can only imagine why they passed up such an opportunity. Goodness-of-fit tests for 14 words in the Shakespeare and Fletcher control texts are presented in Appendix B. Mosteller and Wallace's results are supported; out of 28 tests only one distribution does not fit the

negative binomial. Only two can be adequately described by the Poisson. A number of the observed distributions are extremely non-Poisson, with χ^2 values well over 100. Therefore, the statistical technique proposed by McColly and Weier seems totally inappropriate for the subject of this dissertation, and its value in other authorship questions must be questioned.

Damerau's study "The Use of Function Word Frequencies as Indicators of Style" [24] is another study that uses the Poisson model to study the word occurrence. Mosteller and Wallace and Austin regarded a number of words as "dangerous" because of fears that they were context-dependent. Damerau attempts to develop a more objective way of recognizing context-independent words, but his assumption that a word is independent of context only if its occurrences followed a Poisson distribution is difficult to justify. Mosteller and Wallace demonstrate that, in general, the negative binomial distribution provides a better description of the occurrence of words. Results in Appendix B show that only two of the 14 function words tested has a Poisson distribution in either Shakespeare or Fletcher; by Damerau's definition, almost all of these words are context-dependent. His own results imply that words like *a, and, in* and *of* are context-dependent in samples of Vonnegut, Hemingway, Arthur Miller and Thackeray.

One last authorship study that uses word rates will be reviewed. In a contribution to a general study of the authorship of the Book of Mormon, Larsen and Rencher describe the multivariate statistical techniques they use to analyze what they term "wordprints" [66].[15] While the main emphasis of the study is on the frequency of common non-contextual words, the frequencies of rare words and of letters are also analyzed. Samples of the Book of Mormon and samples of several nineteenth-century writers were analyzed.

---

[15]I am indebted to Dr. Noel Reynolds, the editor of this book (entitled *Book of Mormon Authorship*), for providing me with a photocopy of this section of his publication while he was a visitor at the University of Edinburgh.

Three statistical methods were used in this study: multivariate analysis of variance, cluster analysis and a discriminant analysis. Unfortunately this article is intended for a non-statistical audience and glosses over many important statistical details. It appears that the linear or quadratic discriminant function was employed and that word frequencies were transformed using the arcsine transformation. Distributional assumptions are not discussed or tested. The authors do not closely examine the internal variation of their variables: "We made the same assumption, then, that has been generally accepted and proven widely applicable: each author has a wordprint." Therefore, this paper is difficult to evaluate, but it does make use of modern multivariate techniques that have rarely been applied in authorship studies.

3.4.3 A Syntactic Analysis of Shakespeare and Fletcher

The final study reviewed in this chapter is presented primarily as an indication of the direction in which authorship studies may one day move. In "Authorship Attribution in Jacobean Dramatic Texts" [5] Baillie presents the results of a pilot study of a syntactic analysis of *Henry VIII.* Modern editions of four plays were examined: Shakespeare's *The Winter's Tale* and *Cymbeline,* and Fletcher's *The Woman's Prize* and *Valentinian.* The program EYEBALL was used to label each word in the text according to class (part of speech) and function. The basic data used was ten 500 word samples drawn from each play; these were supplemented with whole scenes and the complete speeches of individual characters. The total number of words processed was around 30,000.

The statistical results that Baillie presents are not impressive. The best individual discriminator he finds is the rate of occurrence of "noun modifiers," which include adjectives, determiners and intensifiers. Baillie gives Shakespeare's mean proportion as 22% (standard deviation 2.3), and Fletcher's as 19% (standard deviation 1.6). He notes that the midpoint between the means correctly assigns 31 of the 40 standard samples, but it seems clear that he does not use statistical significance tests to determine if the means are really different. He proceeds to

test variables in combinations, eventually using the statistical package BMD to perform a multivariate linear discriminant analysis. Although he is satisfied with the seemingly moderate accuracy of these results, the number of samples used in the study (at most 20 for each author) is probably too small.[16]

This study makes no attempt to follow Baillie's approach. Although at the conclusion of his paper he states that an examination of *Henry VIII* is now possible, to my knowledge he has not published the results of such a study. Also, he does not discuss any difficulties encountered in processing his texts with EYEBALL or the number of errors corrected by hand after initial processing. I contacted someone who maintains the similar OXEYE software at a major humanities computing center in England who advised that the program would be rather inaccurate in processing 17th century dramatic texts. Manually correcting the output of an inaccurate program is out of the question for the number of plays used in this study. Recent software developments may have produced a more accurate and sophisticated program than EYEBALL. The CLAWS system used to process the million-word Lancaster-Oslo/Bergen Corpus of British English achieved an accuracy rate of about 97% [79]. Of course, this program would have to be altered for application to Early Modern English before it could be used in Shakespearean problems. But some day the study of the frequency of syntactic classes and structures in large numbers of texts may open many new doors in the area of authorship study.

This chapter demonstrates that researchers in a single area of stylometry have produced a large variety of variables and techniques. A number of these

---

[16]Discriminant analysis will be examined in detail in Chapter 6, where such a procedure is used with word rates to assign authorship for undisputed samples of Shakespeare and Fletcher. While the following comments anticipate many of the findings of that chapter, they will indicate the potential value of tests based on function words. The kernel classification method was used to test the effectiveness of three frequent markers, *the, of* and *in*. In classifying scenes of at least 500 words, 81% of the control set (371 scenes) and 75% of the test set (88 scenes) were assigned to the correct author. These results are comparable to Baillie's. Using a linear discriminant analysis with syntactic data, he achieved a 78% correct classification rate for his design set (40 samples of 500 words) and 78% for his test set (26 complete scenes and 38 character samples).

show promise for the analysis of problems like the question of collaboration in *Henry VIII* and *The Two Noble Kinsmen*. Some of the methods of positional stylometry will be examined next, followed by a study based on word rates.

Chapter 4

Collocations and Proportional Pairs

The previous chapter described some of the variables and procedures of positional stylometry used in authorship studies. This analysis will evaluate the effectiveness of two of these variables, collocations and proportional pairs, in classifying samples of Shakespeare and Fletcher. The large number of variant spellings in old-spelling texts would seem to complicate any definition of a once-occurring word, so tests based on this class are not considered. Likewise tests based on the position of common words in preferred positions in the sentence are not evaluated; punctuation in a given text is inconsistent, and the author's intent was certainly altered in the printing process. For studies of 17th century plays collocations and proportional pairs are the most appropriate of the four types of variables described by Michaelson, Morton and Hamilton-Smith in "To Couple Is the Custom."

The assumptions of stylometry will be evaluated in the control sets of Shakespeare and Fletcher for a number of these variables. The obvious goal is to determine which collocations and proportional pairs occur at very different rates in the works of Fletcher and Shakespeare in comparison to the internal variation within the two control sets. Before differences in rate of use can be evaluated, the extent of contracted forms for the elements of collocations must be measured. Within-author variation is examined between individual plays; in

addition, Shakespeare's plays are grouped according to genre and date of composition to study the effect of these factors on rate of use. Statistical tests of correlation are used to determine if tests based on these features can be treated as if they were independent. Finally, the tests that best discriminate between the two playwrights are evaluated in the test set to determine how successful they can be in identifying samples of unknown or disputed provenance.

4.1 Features Selected for Analysis

The first step in the analysis was to choose a subset of the many individual variables used by Morton, Merriam, Smith and others in earlier studies. Table 4–1 on page 143 provides a list of the thirty collocations chosen from those used in these studies. Most of the collocations used in these studies were considered. A number occur infrequently in Renaissance dramas and were not selected, but the table does not exclude any common collocations.

All collocations examined are "followed by" collocations. In the tables and discussion that follow *and the* represents the proportion of occurrences of *and* that are followed by *the*. Note that collocations of the form "*keyword* followed by *adjective*" have not been tested. Counting adjectives (no matter what definition for an adjective is used) in 34 plays is a completely manual process (at present). Analysis of word combinations by grammatical class may certainly prove useful, but this researcher intends to wait until software is available to make such a study in large bodies of text a reasonable undertaking. Five proportional pairs were also studied; these will be discussed below in Section 4.1.3.

4.1.1 Collocations and their Counts

Several of the collocations listed in Table 4–1 are variants of another collocation. Where the final word of a collocation is the indefinite article *a*, another count has been made which includes occurrences of *an* in the final position. Also, counts of

Keyword	followed by	Shakespeare Ct.	Prop.	Rate	Fletcher Ct.	Prop.	Rate
a		7440			2736		
	X and	402	5.40%	0.95	165	6.03%	1.27
	X of	575	7.73%	1.36	145	5.30%	1.11
and		12398			4301		
	all	152	1.23%	0.36	100	2.33%	0.77
	the	384	3.10%	0.91	85	1.98%	0.65
	X the	532	4.29%	1.26	182	4.23%	1.40
by		1856			419		
	the	312	16.81%	0.74	45	10.74%	0.35
I		11084			4032		
	am	1055	9.52%	2.50	450	11.16%	3.45
	did	151	1.36%	0.36	19	0.47%	0.15
	do	429	3.87%	1.02	124	3.08%	0.95
	have	879	7.93%	2.08	351	8.71%	2.69
in		5557			985		
	a	252	4.53%	0.60	54	5.48%	0.41
	a/an	284	5.11%	0.67	63	6.40%	0.47
	the	832	14.97%	1.97	80	8.02%	0.61
is		5065			790		
	a	318	6.28%	0.75	67	8.48%	0.51
	a/an	370	7.30%	0.88	71	8.99%	0.54
	the	369	7.28%	0.88	55	6.96%	0.42
it		4192			1303		
	is	604	14.41%	1.43	33	2.53%	0.25
	is/was	738	17.60%	1.75	58	4.45%	0.44
of		8325			1655		
	a	298	3.58%	0.71	96	5.80%	0.74
	a/an	341	4.10%	0.81	105	6.34%	0.81
	all	183	2.20%	0.43	79	4.77%	0.61
	the	729	8.76%	1.73	112	6.77%	0.86
	X and	486	5.84%	1.15	107	6.44%	0.82
the		13508			2975		
	X and	793	5.87%	1.88	166	5.58%	1.27
	X the	211	1.56%	0.50	69	2.32%	0.53
	X X the	758	5.61%	1.80	135	4.54%	1.04
to		9388			2552		
	a	163	1.74%	0.39	29	1.14%	0.22
	a/an	182	1.94%	0.43	30	1.17%	0.23
	be	530	5.65%	1.26	96	3.76%	0.74
	the	710	7.56%	1.68	130	5.09%	1.00

Note: "Prop." is the proportion of keywords marked by the collocation. "Rate" is the number of collocations per 1000 tokens.

Table 4–1: Counts and rates for 30 collocations in control plays

the proportion of *it* followed by *is* or *was* are given in conjunction with counts for *it is*. In evaluating the apparent anomaly of the occurrences of *it is* in Fowles' novels, Morton noted that a change of tense was responsible. Combining the two collocations *it is* and *it was* eliminated this apparent difference within Fowles' works [108, p. 144]. If one wished to argue that stylometric habits were due to syntactical characteristics of a writer, then the *is/was* and *a/an* forms would seem more preferable. These distinctions are not critical in the current analysis, as can be observed in the discussion and tables that follow.

The counts shown in Table 4–1 were made from the unexpanded versions of the control texts and therefore do not reflect any occurrences of contracted forms such as *it's* or *i'th'*. Each collocation is listed under its *keyword,* and for each author three numbers are given: the number of occurrences of the collocation; the proportion of keywords marked by the collocation; and, the expected number of occurrences of the feature in a sample of 1000 words. While the proportion of keywords characterized by a given collocation is the variable that is analyzed in the statistical analysis, this third number gives an indication of how often a given collocation might occur in a disputed sample.

Inspection of these figures shows that collocations do not occur as frequently as one might have hoped. In the unexpanded texts the most frequent collocation in both dramatists is *I am,* which occurs at the rate of 2.50 per thousand tokens in Shakespeare and 3.45 in Fletcher. In "To Couple Is the Custom" Morton suggests that the minimum sample size that can be tested must contain at least 5 occurrences of the collocation or proportional pair. For *I am* this minimum size would therefore be about 2000 tokens in Shakespeare and 1450 tokens in Fletcher. Only three scenes in *The Two Noble Kinsmen* and *Henry VIII* are 2000 words or longer. It thus appears that collocations will not be useful in testing individual scenes unless they can be used with samples with fewer than the minimum suggested by Morton.

4.1.2 Contraction and Collocations

Table 4–2 on page 146 contains the counts and rates for the set of collocations after contractions have been expanded (as described in Chapter 2). For most collocations, the change in the proportion of occurrence is small, but the counts for some are seriously altered. The most remarkable difference between the two tables is the dramatic increase in the Fletcher rates of *it is, it is/was* and *in the*. His rate of usage for *it is* shows an increase of over 8 times the rate in the unexpanded texts, and the rate of *in the* almost doubles. In Shakespeare the largest change is an increase of 60% for *it is*. The collocation *it is/was* becomes the most frequent collocation with a rate of 4.10 per 1000 words in Fletcher and 3.55 in Shakespeare. The increase of occurrences of these three frequent collocations brings the low Fletcher rates much closer to the Shakespeare rates. Before proceeding with an analysis of collocations, the "form versus meaning" issue must be confronted once again.

If one observed the uncontracted rates for the tests *it is* or *it is/was* without considering contractions, one would rejoice in the discovery of a test that shows very large differences between samples of Shakespeare and Fletcher. For *it is* the overall proportion in Shakespeare is 14.41% compared to Fletcher's 2.68%, statistically a highly significant difference. When contractions are expanded into their full forms the corresponding values are 23.2% and 21.9%. In other words, 93% of the occurrences of *it is* in Fletcher are bound up in contracted forms, compared to 54% for Shakespeare. While a large difference exists it does not lie in the adjacent use of *it* and *is* but in the rate of contraction of these two words. As shown in Section 2.6, *is* and *it* are two of the words that most often appear in contracted forms. Moreover the two authors' rates of contraction for *is* are significantly different, although date of composition affects this rate in Shakespeare. The difference in the frequencies of *it is* is a side-effect of this more general characteristic.

The contracted forms of *it is* have been treated in varying ways in the previous studies of collocations in Jacobean drama. As noted earlier, Morton did

Keyword followed by		Shakespeare Ct.	Prop.	Rate	Fletcher Ct.	Prop.	Rate
a		7440			2736		
	X and	403	5.42%	0.95	165	6.03%	1.26
	X of	587	7.89%	1.38	149	5.45%	1.14
and		12416			4328		
	all	152	1.22%	0.36	101	2.33%	0.77
	the	384	3.09%	0.90	85	1.96%	0.65
	X the	527	4.25%	1.24	181	4.18%	1.38
by		1874			427		
	the	326	17.40%	0.76	48	11.24%	0.37
I		12033			4518		
	am	1059	8.80%	2.48	464	10.27%	3.54
	did	151	1.26%	0.35	19	0.42%	0.14
	do	429	3.57%	1.01	124	2.75%	0.95
	have	879	7.31%	2.06	351	7.77%	2.68
in		5763			1141		
	a	252	4.37%	0.59	54	4.73%	0.41
	a/an	284	4.93%	0.67	62	5.43%	0.47
	the	979	16.99%	2.30	170	14.90%	1.30
is		6971			2029		
	a	526	7.55%	1.23	192	9.46%	1.46
	a/an	603	8.65%	1.41	205	10.10%	1.56
	the	506	7.26%	1.19	148	7.29%	1.13
it		5610			2172		
	is	1303	23.23%	3.06	473	21.77%	3.61
	is/was	1514	26.99%	3.55	541	24.91%	4.13
of		8489			1718		
	a	298	3.51%	0.70	96	5.59%	0.73
	a/an	341	4.02%	0.80	105	6.11%	0.80
	all	183	2.16%	0.43	79	4.60%	0.80
	the	864	10.18%	2.03	175	10.19%	1.33
	X and	486	5.73%	1.14	107	6.23%	0.82
the		13873			3149		
	X and	822	5.93%	1.93	183	5.81%	1.40
	X the	217	1.56%	0.51	74	2.35%	0.56
	X X the	836	6.03%	1.96	149	4.73%	1.14
to		9512			2583		
	a	163	1.71%	0.38	29	1.12%	0.22
	a/an	182	1.91%	0.43	30	1.16%	0.23
	be	530	5.57%	1.24	96	3.72%	0.73
	the	773	8.13%	1.81	145	5.61%	1.11

Note: "Prop." is the proportion of keywords marked by the collocation. "Rate" is the number of collocations per 1000 tokens.

Table 4–2: Counts and rates for 30 collocations after contractions are expanded

not count contracted forms in studying *Pericles,* and Smith discovered that this choice had some effect on the results of a collocation-based study of the play. Merriam included *'tis* and *it's* as occurrences of the collocation in his study of *Henry VIII* [89], but in "The Consonance of Literary Elements with Mathematical Models" he argues against the expansion of contracted forms [92, pp. 231–233].

Noting that computer software must be able to recognize all contracted forms, Merriam warns that automatic processing may not result in consistent results. However this is true even for a machine count of simple words in 17th century texts. As shown in Section 2.5.3 (page 60), some degree of standardization is required in order to produce accurate counts of something as common as the indefinite article. Merriam also notes the possible effects of compositors and scribes but does not view this as support for altering the orthographical forms found in a text. He remarks that "any consistent method of counting habits is valid" and relies on the redundancy (in the information theory sense of the word) in stylometry to allow for any effect resulting from ignoring contracted forms. This approach assumes that different counting conventions will not change the results of a statistical analysis. Such an approach may be valid in situations where authors do not exhibit widely-differing practices, but Section 2.6 shows that this is not the case for some words in Shakespeare and Fletcher.

Earlier in the same work, Merriam compares samples from *The Two Noble Kinsmen* with two plays by Shakespeare and two by Fletcher as an illustration of the effectiveness of collocation tests. Comparison with my own data confirms that his counts do not include any contracted forms. The tables he presents on pages 140–141 show that in *Monsieur Thomas* and *Valentinian* the word *it* occurs 462 times, of which 16 occurrences (3.5%) are followed by *is.* My counts show that, if the contracted forms are expanded to their full forms, the numbers increase to 676 and 147 (21.8%).

This single test is crucial to the results Merriam presents. Using eight collocations he compares Act I (plus III.i) of *The Two Noble Kinsmen* to the Shakespeare and Fletcher texts, finding a 200 to 1 likelihood ratio in favor of Shakespeare's authorship. If the collocation *it is* is excluded from the comparison, the remaining seven tests result in a ratio of under 5 to 1 (according to the method of multiplying probabilities used). Another example attempts to classify a single scene of under 500 words, II.ii. The eight collocations result in a likelihood ratio of 10 to 1 in favor of Shakespeare, a result which Merriam considers "provisional" due to the small sample size. (In fact there are only 12 occurrences of any of the eight collocations in the entire scene.) Excluding *it is* from the comparison produces an even less satisfactory result: a 1.2 to 1 likelihood ratio in favor of Fletcher.

To attach great significance to the different rates of *it is* in the unexpanded texts of Shakespeare and Fletcher is to neglect the more fundamental difference in the rate of contraction. Section 2.2.2 (page 31) describes the evidence that distinctions between contracted and full forms were not always preserved in the transcription or printing process. In addition, the results in Table 2–5 (page 76) indicate that Shakespeare's use of some contracted forms is strongly related to a play's date of composition. If the dates of the disputed sample were unknown, then the relationship of contraction rates to authorship would be difficult to interpret.

For these reasons I maintain that stylometric studies in Jacobean drama should be based on the expanded forms of compound contractions. Expansion of contractions helps move stylometry from an orthographical to a lexical basis. The nature of dramatic dialogue and the possibility of textual modification in transmission support the principle that a method for determining authorship should be robust regarding differences of surface orthographical features. Therefore, in this dissertation, discussion of the frequencies of single words, collocations and proportional pairs will focus on counts from the expanded versions of the Shakespearean and Fletcher texts.

Prop. Pair		Shakespeare Ct.	Prop.	Rate	Fletcher Ct.	Prop.	Rate
an+a		8274			2990		
	an	834	10.08%	1.96	254	8.49%	1.91
any+all		2336			1212		
	any	480	20.55%	1.13	191	15.76%	1.43
no+not		6515			2171		
	no	1966	30.18%	4.61	844	38.88%	6.34
this+that		9107			2724		
	this	3319	36.44%	7.79	1144	42.00%	8.59
without+with		3899			955		
	without	202	5.18%	0.47	70	7.33%	0.53

Note: "Prop." is the count for the first member of the pair divided by the total count for both. "Rate" is the first member's rate per 1000 tokens.

Table 4–3: Counts and rates for 5 proportional pairs in expanded control texts

4.1.3 Proportional Pairs

The five proportional pairs examined are listed in Table 4–3 with the overall counts, proportions and frequency in both Shakespeare and Fletcher. A large number of proportional pairs have been used in the studies described in Section 3.3. Only a limited number of frequent collocations can be found in English texts, but any two words can be tested as a proportional pair. The five used in this study were chosen from among those listed in *Literary Detection* and "The Nature of Stylometry" [112]. Table 4–3 shows that several occur more often than collocations, especially *no+not* and *this+that*. (A word about notation: in the discussion that follows "*no+not*" is shorthand for "the number of the first word, *no*, in proportion to the total count of the pair.")

4.2 Measuring Differences between Authors

The basic premise of a stylometric authorship study is that habits exist which vary significantly between authors and little within authors' works. A feature must meet both these criteria in order to be useful in a study of authorship. The authors of "To Couple Is the Custom" begin by testing collocations and proportional pairs for consistency within the works of Scott, James and Fowles. I am reversing the process in this study; after finding the features that best discriminate between Shakespeare and Fletcher, I will compare internal variation to the magnitude of the differences between the two.

4.2.1 χ^2 tests

The χ^2 test has often been used to determine whether a given feature (such as a collocation or a proportional pair) is a "habit" for two authors that can also be used to discriminate between their works. One common characteristic of almost all the positional stylometry studies described in Chapter 3 is the use of tests (usually with the χ^2 test) on data arranged in contingency tables to determine if two writers differ. In this situation, the χ^2 test evaluates the null hypothesis that data classified by sample and occurrence of a feature shows no association between classifications. In other words, the number of occurrences of a feature will be distributed in the samples according to the overall proportion. When the samples are works known to be by different authors, a high χ^2 value is considered to reflect the difference in authorship.

Results of tests on contingency tables are the basis for the classification of unknown samples in the studies such as "To Couple Is the Custom" and Smith's study of *Pericles* [144]. The data for *by the* in Shakespeare and Fletcher listed in Table 4–2 can be used to illustrate this use of χ^2. The counts are arranged in a 2×2 contingency table:

Author	*by* FB *the*	*by* NOT FB *the*	Total num. of *by*s
Fletcher	48	380	428
Shakespeare	326	1548	1874
Totals	374	1928	2302

and a χ^2 value (with or without Yates' correction) is calculated. As noted in Section 3.3.4 Merriam advocates the use of Fisher's exact probability test over χ^2 tests in analyzing 2×2 tables. The exact test is more accurate, especially for tables with small numbers, but is more difficult to calculate without computer software than χ^2.

While the computation of a χ^2 test is based on the actual number of occurrences of the keyword and the "mark" word, one can divide the two counts and use this proportion (a single continuous variable) in a number of other statistical calculations (for example, the analysis of variance). There are some subtleties to this change of variables. For example, an entire play may have the same proportion of occurrence for a collocation as a small scene containing far fewer occurrences of the keyword. When using a 2×2 contingency table to compare either one to a third sample, the χ^2 test allows for the the different number of keyword occurrences in the samples. Any method operating directly on the proportion alone can only recognize the two measurements as identical. Certainly this is one advantage of tests based on contingency tables. One should also note that the statistical distribution of a features's counts in equal-length samples may be different from the distribution of the same feature's rate of occurrence in samples with different lengths.

4.2.2 *t* Tests

A χ^2 test (or exact probability test) that compares the overall counts of a feature for two writers makes no allowance for the observed variance within each writer's samples. A collocation's proportion of occurrence will be used in another statistical test to compare the mean rates in the two dramatists. If a feature is

counted in a number of samples then t tests can be used to determine whether the mean proportions for each writer are significantly different. This method evaluates the difference between the means $\overline{x}_1 - \overline{x}_2$ according to the variance of this difference: $\sigma^2_{\overline{x}_1 - \overline{x}_2} = \sigma_1^2 + \sigma_2^2$. The resulting statistic:

$$t' = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \tag{4.1}$$

unfortunately does not follow Student's t distribution when the two variances are unequal, but Snedecor and Cochran [151, p. 97] and Bailey [4, p. 51] each give an approximation which assigns a number of degrees of freedom ν' to t' so the ordinary t table can be used. In this study the former approximation (due to Satterthwaite) will be used:

$$\nu' = \frac{(v_1 + v_2)^2}{(v_1^2/\nu_1 + v_2^2/\nu_2)} \tag{4.2}$$

where $v_i = s_i^2/n_i$ and $\nu_i = n_i - 1$. Generally this value ν' is not an integer and should be rounded down before making use of a standard table for t.

4.2.3 Features that Discriminate

Three methods were used to compare the rate of occurrence for the set of collocations. χ^2 with Yates' correction and Fisher's exact probability test were computed for the overall counts as listed in Table 4–2. The statistic t' of Equation 4.1 was also calculated for the difference between Fletcher's and Shakespeare's means. (Thus words favored by Fletcher have negative values of t'.) For the exact test and the χ^2 test the probability that the differences observed are due to random variation is given. Likewise the table includes the t test probability that the mean rates of occurrence in the two authors are the same. The values of t' were calculated from the counts of collocations in the 101 acts of the Shakespeare control set and the 30 acts of the Fletcher. (The mean size of an act in the expanded texts of the Shakespeare control set is 4420 words, with a standard deviation of 1394 words; the smallest act in the set is 1511 words and the largest 8700. For the Fletcher set the mean is 4439 words with a standard

deviation of 1139; the smallest is 2295 words and the largest 7783.) The results of all three tests can be compared in Table 4–4 on page 154; the entries are sorted by decreasing value of the probability associated with t'.[1]

Not surprisingly the results for the χ^2 and exact probability tests are very close for these large samples. The probabilities derived from the t test agree for the most part with the other two probabilities. There are some interesting exceptions. The data for *to a* yields a small t' probability indicating a very significant difference between the Shakespeare and Fletcher means, but the probabilities resulting from the χ^2 and exact probability tests are borderline at the 5% level. This collocation is infrequent and has a small sample variance in both writers in comparison with the difference in means, which produces the significant t' value. The opposite is true of *I am*. The χ^2 and exact probability tests show a very large between-author difference, but this is small compared to the variance within the writers. The probability associated with a t of 2.04 with 50 degrees of freedom is only just under the 5% level.

The discussion of the previous paragraph demonstrates that the separation of the two criteria of "internal consistency" and "discrimination between writers" is artificial. Measurements of one do not mean anything until compared to the other. Before examining the within-writer variation several conclusions can be drawn from Table 4–4. The two writers use some collocations at different rates: the more frequent of these are *to the, a* X *of* and *the* X X *the*. Others are less frequent but possibly better discriminators: *I did, of all* and *and all*. The most frequent collocations (such as *it is* and *I am*) fall in the bottom half of the table and would appear to be less useful than those already mentioned. Two frequent

---

[1]The χ^2 values and the probabilities for the exact test were calculated by Pascal programs I wrote. I am very grateful to my supervisor, Professor Sidney Michaelson, for spending numerous hours writing and testing an IMP procedure that calculates the associated probability for a given χ^2 value. (I have translated this procedure into PASCAL and FORTRAN.) t' and the degrees of freedom ν' were computed by a small program written in awk on a computer running the UNIX operating system. The associated probabilities for these statistics were then computed using a Fortran function in the NAG mathematical library on Edinburgh's EMAS system.

Collocation or Prop. Pair	χ^2 with Yates' value	prob.	Exact text prob.	t test t'	ν'	prob.
no+not	55.915	0.0000	0.0000	4.83	43.72	0.0000
I did	21.689	0.0000	0.0000	-5.96	123.53	0.0000
of all	32.902	0.0000	0.0000	4.24	38.70	0.0001
to the	18.820	0.0000	0.0000	-4.31	66.55	0.0001
and all	26.484	0.0000	0.0000	3.76	41.57	0.0005
to a/an	5.794	0.0161	0.0115	-3.22	83.28	0.0018
a X of	17.796	0.0000	0.0000	-3.21	58.16	0.0022
of a	16.778	0.0000	0.0001	3.26	36.95	0.0024
this+that	27.282	0.0000	0.0000	3.16	43.14	0.0029
of a/an	15.136	0.0001	0.0002	2.84	36.72	0.0074
to a	3.778	0.0519	0.0422	-2.73	82.41	0.0077
to be	14.502	0.0001	0.0001	-2.71	51.76	0.0091
an+a	6.141	0.0132	0.0116	-2.65	62.88	0.0101
I do	6.609	0.0101	0.0087	-2.59	90.77	0.0112
the X X the	7.715	0.0055	0.0047	-2.60	61.21	0.0117
by the	9.334	0.0022	0.0014	-2.57	59.23	0.0127
is a	7.342	0.0067	0.0069	2.47	44.91	0.0175
and the	14.855	0.0001	0.0001	-2.43	49.85	0.0188
any+all	11.623	0.0007	0.0005	-2.12	71.88	0.0379
the X the	8.653	0.0033	0.0038	2.15	37.97	0.0382
without+with	6.297	0.0121	0.0119	2.08	42.51	0.0437
I am	8.287	0.0040	0.0041	2.04	50.91	0.0467
is a/an	3.687	0.0548	0.0523	1.88	47.85	0.0663
in the	2.911	0.0880	0.0847	-1.70	46.28	0.0959
the X and	0.024	0.8772	0.8632	-0.69	43.75	0.4939
is the	0.001	0.9698	0.9991	-0.64	59.19	0.5247
it is/was	3.018	0.0824	0.0817	-0.64	53.92	0.5249
of the	0.001	0.9804	1.0000	0.59	57.90	0.5575
and X the	0.031	0.8605	0.8595	-0.58	57.77	0.5642
a X and	1.225	0.2683	0.2632	0.49	55.98	0.6261
I have	0.955	0.3286	0.3222	0.48	43.41	0.6337
it is	1.586	0.2079	0.2070	-0.37	57.28	0.7128
in a	0.296	0.5867	0.5306	-0.29	42.28	0.7732
in a/an	0.508	0.4760	0.4168	-0.09	39.2	0.9287
of X and	0.539	0.4629	0.4305	0.02	40.47	0.9841

Table 4–4: Between-author comparison using χ^2, exact and t tests

proportional pairs, *no+not* and *this+that,* also show promise for distinguishing Fletcher from Shakespeare. But before using any of these features to assign disputed samples, the internal variation within the works by Shakespeare and Fletcher must be analyzed.

4.3 Internal Variation

χ^2 tests have also been used extensively in stylometric studies to demonstrate consistency within a writer's work. Often the problem itself is one of homogeneity. In any problem samples of known authorship are examined to ensure that an author is consistent in his usage. χ^2 has almost always been used to show that it would be highly unlikely that the samples came from different populations. For example, Morton and his colleagues used χ^2 tests to show that the tests that discriminated between Jane Austen and the "Other Lady" showed homogeneity in *Sense and Sensibility, Emma* and the authentic parts of *Sanditon.* O'Brien and Darnell divided known samples into two halves and tested one against the other.

To test the hypothesis that collocations and proportional pairs are used consistently within the works of Shakespeare and Fletcher, counts were made for each feature in each of the control plays. For each author a $n \times 2$ contingency table was evaluated, where n (the number of plays) is 20 for Shakespeare and 6 for Fletcher. The probabilities associated with the χ^2 value for each feature are given in Table 4–5.

These χ^2 results are not encouraging. When the 20 Shakespeare plays are tested, 27 of the 30 collocations have χ^2 values that are significant at the 5% level; 12 of 30 are significant for the 6 Fletcher texts. Using the 1% level, the number of significant tests decreases to 24 for Shakespeare and 5 for Fletcher. Only 3 collocations are not significant at the 5% level for both writers: *a* X *and, in a* and *in a/an.* Only 3 others have probabilities for both between the 5% and 1% levels: *and all, in the* and *the* X *and.* The proportional pairs fair somewhat

Collocation or Prop. pair	Shakespeare χ^2	prob.		Fletcher χ^2	prob.	
a X and	27.92	0.0850	*	7.06	0.2164	*
a X of	40.78	0.0026		24.04	0.0002	
and all	33.17	0.0230	*	2.24	0.8143	*
and the	50.72	0.0001		9.12	0.1043	*
and X the	39.11	0.0043		2.95	0.7077	*
by the	42.05	0.0017		11.11	0.0492	*
I am	38.96	0.0045		10.83	0.0549	*
I did	36.63	0.0088		2.61	0.7595	*
I do	43.20	0.0012		16.15	0.0064	
I have	64.56	0.0000		8.75	0.1195	*
in a	22.09	0.2796	*	5.03	0.4127	*
in a/an	21.83	0.2928	*	3.71	0.5921	*
in the	34.20	0.0174	*	12.80	0.0253	*
is a	38.89	0.0046		13.21	0.0214	*
is a/an	52.77	0.0001		15.27	0.0092	
is the	40.72	0.0026		4.62	0.4642	*
it is	49.68	0.0001		24.19	0.0002	
it is/was	45.87	0.0005		21.93	0.0005	
of a	40.09	0.0032		2.02	0.8459	*
of a/an	41.94	0.0018		2.34	0.8003	*
of all	56.53	0.0000		4.31	0.5057	*
of the	118.62	0.0000		8.21	0.1451	*
of X and	40.25	0.0030		4.27	0.5106	*
the X and	31.35	0.0369	*	11.67	0.0396	*
the X the	72.96	0.0000		11.71	0.0389	*
the X X the	56.09	0.0000		5.13	0.4006	*
to a	50.18	0.0001		2.33	0.8015	*
to a/an	49.98	0.0001		2.62	0.7584	*
to be	47.37	0.0003		12.47	0.0289	*
to the	54.15	0.0000		13.19	0.0217	*
an+a	21.40	0.3150	*	4.44	0.4874	*
any+all	111.84	0.0000		15.50	0.0084	
no+not	18.18	0.5103	*	22.58	0.0004	
this+that	83.29	0.0000		28.91	0.0000	
without+with	24.84	0.1659	*	8.25	0.1432	*

Note: For tests of the Shakespeare control tests, degrees of freedom = 19. For the Fletcher control, degrees of freedom = 5. Tests *not* signficant at the 1% level are marked with an asterisk.

Table 4–5: χ^2 tests for internal consistency

better. In Shakespeare 3 of the 5 are not significant at the 5% level, and two of these are also non-significant in the Fletcher plays.

Referring back to Table 4–4 one notes that, of the six collocations that do pass the χ^2 criteria for homogeneity at the 1% level, only *and all* shows a significant difference in rate of occurrence between authors. In addition one proportional pair, *an+a,* occurs very consistently within the plays and also has a t' probability of 0.0101. Most of the collocation and proportional pair tests that show large between-author differences are also characterized by large within-author differences in at least one of the two dramatists' control samples.

One conclusion that can be drawn from the data presented in this table is that χ^2 tests are not a reasonable method for studying the pattern of occurrence within a writer's works. In their examination of chapters of Scott's *The Antiquary,* Morton and his associates found that the occurrences of four tests that yielded significant χ^2 values fit the Poisson distribution. Furthermore Merriam has shown that the occurrences of 11 collocations in Shakespeare's First Folio are described by the negative binomial (or the Poisson in one instance) distribution. As noted earlier, these two results raise one's suspicions that the binomial distribution may not be a good model for occurrences of relatively rare literary features (like collocations and proportional pairs) when a large number of samples are examined.

Several studies have proposed a mixture of Poisson distributions as a likely model for certain features of composition. Kemp provides a clear discussion of this idea [59]. Mosteller and Wallace, who successfully utilized the negative binomial, interpret their model as a mixture of Poisson probabilities with a gamma distribution for the Poisson mean [113, pp. 94–95]. Sichel proposes a more generalized (and complex) distribution, of which the Poisson, negative binomial and geometric distributions (among others) are special or limiting forms [141]. He shows that one form of this distribution adequately describes almost all published sentence-length distributions. Such models, based around the intuitively

appealing idea of a basic Poisson model with the mean varying with style or subject, may indeed best describe the distributions of collocations and proportional pairs. If so, the greater variation inherent in these models is probably at the root of the the large χ^2 values in Table 4–5 for the Shakespeare plays.

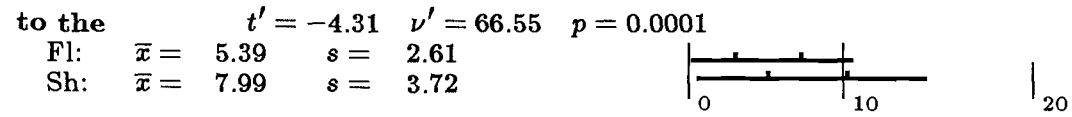
In examining the counts that formed the contingency tables used in these tests, I observed that there was large range of values within each writer's works for most of these features. The large χ^2 values did not appear to be due to a few outliers (observations very distant from the mean). For almost every collocation and proportional pair, there was considerable overlap in values, even if the mean rates of occurrence were significantly different. Rather than illustrate this by printing 35 contingency tables for each author, I have chosen to use graphs to represent the proportion of occurrence in samples (expressed as a percentage), along with the standard deviations and the interquartile range.

For data distributed according to the normal distribution, 95% percent of the observations are expected to lie within two standard deviations of the mean. Even if this data is not generally normal, this range of four standard deviations does provide some indication of the extent of the overlap of occurrence.[2] If a distribution is skew or contains a few outliers then the standard deviation can be somewhat misleading. In such cases one considers the *interquartile range,* the distance between the the first and third quartiles. By definition one half of the observed values lie between these points.

---

[2]Given the low means and noticeable skewness for most of the distributions for these features, it might seem likely that few if any would be distributed normally. However goodness-of-fit tests at the 5% level of significance show that, of the 30 collocations, 15 are distributed normally in acts of Fletcher; 19 of the tests are normal in Shakespeare's acts. Statisticians have developed several goodness-of-fit tests for the normal distribution, and two are used by the SAS statistical package's UNIVARIATE procedure, which calculates simple statistics for univariate data [123]. The Shapiro-Wilk statistic W is based on an analysis of variance and is suitable for distributions with 50 or fewer observations. Kolmogorov's D can be used when the number of observations is larger. Thus the former test was used for the Fletcher observations and the latter for Shakespeare. Note that these tests are for the proportion of occurrence (a continuous variable) in samples of varying length. Merriam's results (described earlier) are for *counts* in samples of the same length.

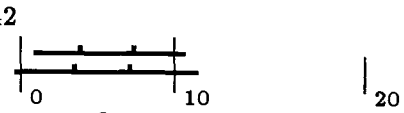
Table 4–6 presents the information for all 35 collocations and proportional pairs. To help one interpret the information in this table, one entry will be described in detail. The following represents data for the collocation *to the* counted in the expanded versions of the texts:



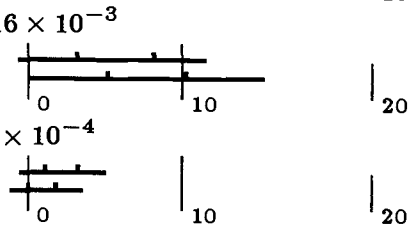
The statistics presented are for the rate of occurrence, measured in each of the 131 acts of the Shakespeare and Fletcher control set. The t test statistics from Table 4–4 are again printed to the right of the collocation identifier. For each author the mean $\overline{x}$ is printed followed by the sample standard deviation s. To the right of these values a bar of length $4s$ is centered around the mean for each author. Thin vertical struts that pass through both horizontal bars indicate the scale of the graph. Two small ticks sitting on top of each of the horizontal bars mark the first and third quartiles. One half of the observed values lie between these marks, and the distance between them is the interquartile range.

Note: The bars on the right indicate $\overline{x} \pm 2s$. The ticks on top of the bars mark the 1st and 3rd quartiles. The statistics are for rate of occurrence, expressed as a percentage.

a X **and** $t' = 0.49$ $\nu' = 55.98$ $p = 0.5642$
Fl: $\overline{x} =$ 5.81 $s =$ 2.46
Sh: $\overline{x} =$ 5.54 $s =$ 2.98



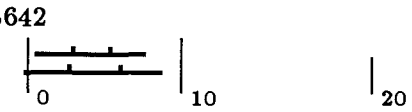
a X **of** $t' = -3.21$ $\nu' = 58.16$ $p = 2.16 \times 10^{-3}$
Fl: $\overline{x} =$ 5.44 $s =$ 3.05
Sh: $\overline{x} =$ 7.62 $s =$ 3.84



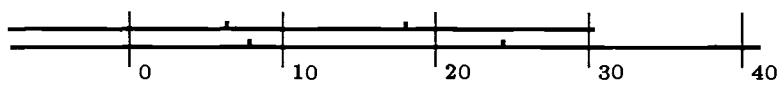
and all $t' = 3.76$ $\nu' = 41.57$ $p = 5.31 \times 10^{-4}$
Fl: $\overline{x} =$ 2.25 $s =$ 1.41
Sh: $\overline{x} =$ 1.19 $s =$ 1.18

and the $t' = -2.43$ $\nu' = 49.85$ $p = 0.0188$
Fl: $\overline{x} =$ 2.10 $s =$ 2.00
Sh: $\overline{x} =$ 3.12 $s =$ 2.14

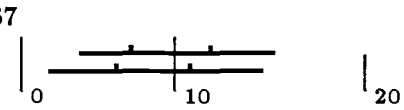
and X **the** $t' = -0.58$ $\nu' = 57.77$ $p = 0.5642$
Fl: $\overline{x} =$ 4.04 $s =$ 1.80
Sh: $\overline{x} =$ 4.27 $s =$ 2.25



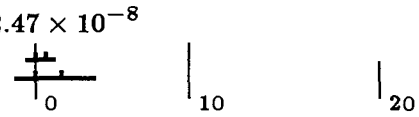
by the $t' = -2.57$ $\nu' = 59.23$ $p = 0.0127$
Fl: $\overline{x} =$ 11.27 $s =$ 9.55
Sh: $\overline{x} =$ 16.74 $s =$ 12.22



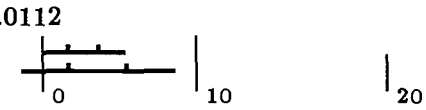
I am $t' = 2.04$ $\nu' = 50.91$ $p = 0.0467$
Fl: $\overline{x} =$ 10.15 $s =$ 3.18
Sh: $\overline{x} =$ 8.77 $s =$ 3.49



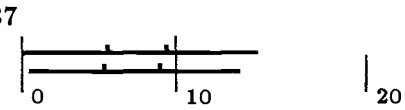
I did $t' = -5.96$ $\nu' = 123.53$ $p = 2.47 \times 10^{-8}$
Fl: $\overline{x} =$ 0.35 $s =$ 0.48
Sh: $\overline{x} =$ 1.29 $s =$ 1.31



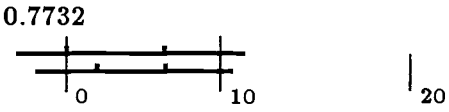
I do $t' = -2.59$ $\nu' = 90.77$ $p = 0.0112$
Fl: $\overline{x} =$ 2.70 $s =$ 1.33
Sh: $\overline{x} =$ 3.60 $s =$ 2.49



I have $t' = 0.48$ $\nu' = 43.41$ $p = 0.6637$
Fl: $\overline{x} =$ 7.66 $s =$ 3.81
Sh: $\overline{x} =$ 7.29 $s =$ 3.42



in a $t' = -0.29$ $\nu' = 42.28$ $p = 0.7732$
Fl: $\overline{x} =$ 4.18 $s =$ 3.71
Sh: $\overline{x} =$ 4.40 $s =$ 3.20



in a/an $t' = -0.09$ $\nu' = 39.22$ $p = 0.9287$
Fl: $\overline{x} =$ 4.92 $s =$ 4.41
Sh: $\overline{x} =$ 5.00 $s =$ 3.35

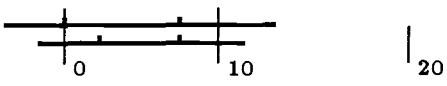
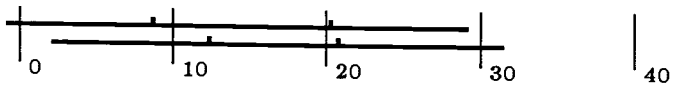


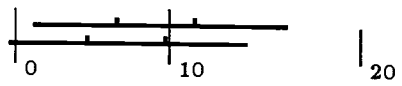
Table 4–6: Internal variation of collocations and proportional pairs

Note: The bars on the right indicate $\overline{x} \pm 2s$. The ticks on top of the bars mark the 1st and 3rd quartiles. The statistics are for rate of occurrence, expressed as a percentage.

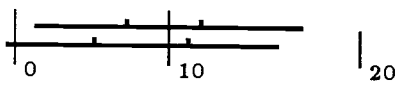
in the $t' = -1.70$ $\nu' = 46.28$ $p = 0.0959$
Fl: $\overline{x} =$ 14.16 $s =$ 7.51
Sh: $\overline{x} =$ 16.79 $s =$ 7.35



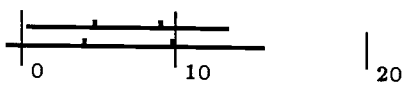
is a $t' = 2.47$ $\nu' = 44.91$ $p = 0.0175$
Fl: $\overline{x} =$ 9.38 $s =$ 4.11
Sh: $\overline{x} =$ 7.30 $s =$ 3.86



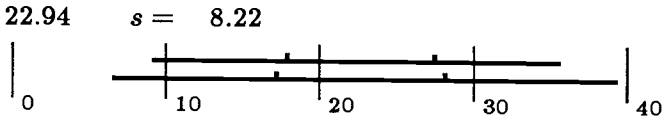
is a/an $t' = 1.88$ $\nu' = 47.85$ $p = 0.0663$
Fl: $\overline{x} =$ 9.97 $s =$ 4.34
Sh: $\overline{x} =$ 8.26 $s =$ 4.42



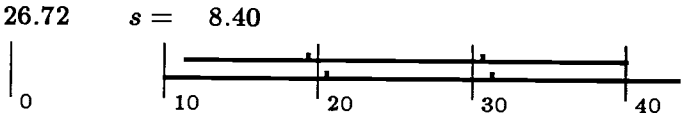
is the $t' = -0.64$ $\nu' = 59.19$ $p = 0.5247$
Fl: $\overline{x} =$ 6.89 $s =$ 3.27
Sh: $\overline{x} =$ 7.36 $s =$ 4.19



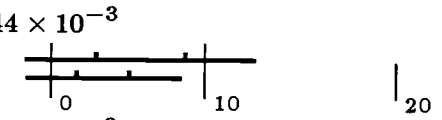
it is $t' = -0.37$ $\nu' = 57.28$ $p = 0.7128$
Fl: $\overline{x} =$ 22.39 $s =$ 6.63
Sh: $\overline{x} =$ 22.94 $s =$ 8.22



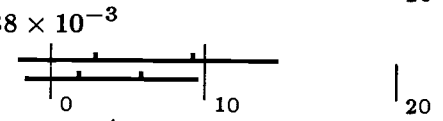
it is/was $t' = -0.64$ $\nu' = 53.92$ $p = 0.5249$
Fl: $\overline{x} =$ 25.72 $s =$ 7.21
Sh: $\overline{x} =$ 26.72 $s =$ 8.40



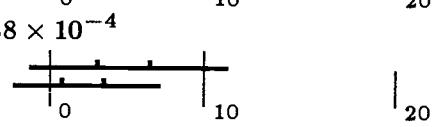
of a $t' = 3.26$ $\nu' = 36.95$ $p = 2.44 \times 10^{-3}$
Fl: $\overline{x} =$ 5.79 $s =$ 3.75
Sh: $\overline{x} =$ 3.42 $s =$ 2.53



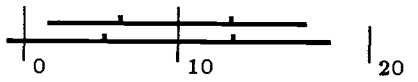
of a/an $t' = 2.84$ $\nu' = 36.72$ $p = 7.38 \times 10^{-3}$
Fl: $\overline{x} =$ 6.30 $s =$ 4.22
Sh: $\overline{x} =$ 3.97 $s =$ 2.81



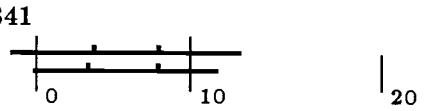
of all $t' = 4.24$ $\nu' = 38.70$ $p = 1.38 \times 10^{-4}$
Fl: $\overline{x} =$ 5.10 $s =$ 3.22
Sh: $\overline{x} =$ 2.42 $s =$ 2.39



of the $t' = 0.59$ $\nu' = 57.90$ $p = 0.5575$
Fl: $\overline{x} =$ 9.91 $s =$ 4.19
Sh: $\overline{x} =$ 9.36 $s =$ 5.24



of X and $t' = 0.02$ $\nu' = 40.47$ $p = 0.9841$
Fl: $\overline{x} =$ 5.79 $s =$ 3.73
Sh: $\overline{x} =$ 5.78 $s =$ 3.00



the X and $t' = -0.69$ $\nu' = 43.75$ $p = 0.4939$
Fl: $\overline{x} =$ 5.68 $s =$ 2.57
Sh: $\overline{x} =$ 6.05 $s =$ 2.33

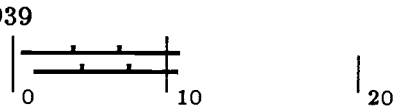
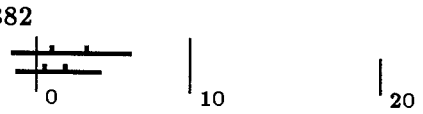


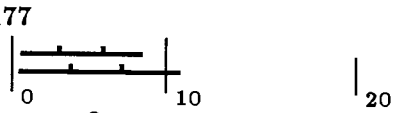
Table 4–6 (*cont.*)**:** Internal variation of collocations and proportional pairs

Note: The bars on the right indicate $\overline{x} \pm 2s$. The ticks on top of the bars mark the 1st and 3rd quartiles. The statistics are for rate of occurrence, expressed as a percentage.

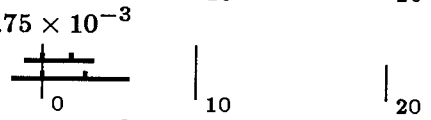
the X the $t' = 2.15$ $\nu' = 37.87$ $p = 0.0382$
Fl: $\overline{x} =$ 2.26 $s =$ 1.95
Sh: $\overline{x} =$ 1.44 $s =$ 1.39



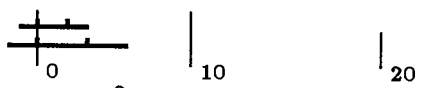
the X X the $t' = -2.60$ $\nu' = 61.21$ $p = 0.0177$
Fl: $\overline{x} =$ 4.50 $s =$ 1.98
Sh: $\overline{x} =$ 5.66 $s =$ 2.62



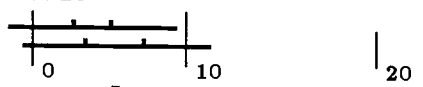
to a $t' = -2.73$ $\nu' = 82.41$ $p = 7.75 \times 10^{-3}$
Fl: $\overline{x} =$ 1.07 $s =$ 1.12
Sh: $\overline{x} =$ 1.83 $s =$ 1.92



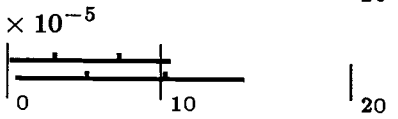
to a/an $t' = -3.22$ $\nu' = 83.28$ $p = 1.83 \times 10^{-3}$
Fl: $\overline{x} =$ 1.09 $s =$ 1.13
Sh: $\overline{x} =$ 2.00 $s =$ 1.96



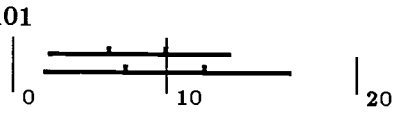
to be $t' = -2.71$ $\nu' = 51.76$ $p = 9.14 \times 10^{-2}$
Fl: $\overline{x} =$ 3.88 $s =$ 2.74
Sh: $\overline{x} =$ 5.47 $s =$ 3.05



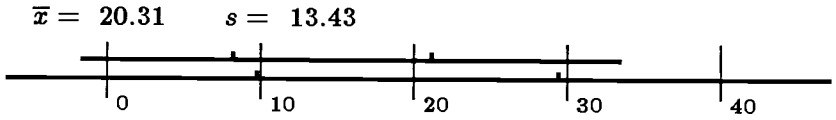
to the $t' = -4.31$ $\nu' = 66.55$ $p = 5.56 \times 10^{-5}$
Fl: $\overline{x} =$ 5.39 $s =$ 2.61
Sh: $\overline{x} =$ 7.99 $s =$ 3.72



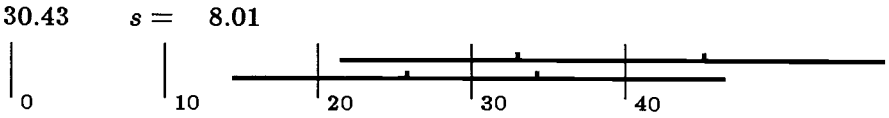
an+a $t' = -2.65$ $\nu' = 62.87$ $p = 0.0101$
Fl: $\overline{x} =$ 8.23 $s =$ 2.97
Sh: $\overline{x} =$ 10.02 $s =$ 4.02



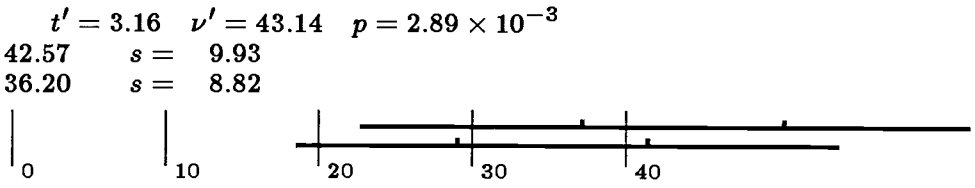
any+all $t' = -2.12$ $\nu' = 71.88$ $p = 0.0379$
Fl: $\overline{x} =$ 15.89 $s =$ 8.80
Sh: $\overline{x} =$ 20.31 $s =$ 13.43



no+not $t' = 4.83$ $\nu' = 43.73$ $p = 1.76 \times 10^{-5}$
Fl: $\overline{x} =$ 39.12 $s =$ 8.84
Sh: $\overline{x} =$ 30.43 $s =$ 8.01



this+that $t' = 3.16$ $\nu' = 43.14$ $p = 2.89 \times 10^{-3}$
Fl: $\overline{x} =$ 42.57 $s =$ 9.93
Sh: $\overline{x} =$ 36.20 $s =$ 8.82



without+with $t' = 2.08$ $\nu' = 42.51$ $p = 0.0437$
Fl: $\overline{x} =$ 7.17 $s =$ 4.78
Sh: $\overline{x} =$ 5.16 $s =$ 4.16

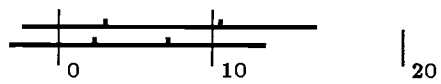


Table 4–6 (*cont.*)**:** Internal variation of collocations and proportional pairs

Inspection of the entries in this table shows that the ranges of occurrences of these tests overlap considerably, even where the t test shows that the means are significantly different. Of the top tests of Table 4–4 (page 154), *of all, and all* and *no+not* appear to overlap the least. Others in this group (such as *to a*) have almost identical quartiles for both authors. For many of the tests it is clear that a Fletcher sample could be as close to that author's mean rate as one half of the total number of Fletcher samples, yet still be much closer to the overall Shakespeare rate. By and large these figures suggest that most collocation rates in acts of Shakespeare and Fletcher are too closely intermingled to be of much use individually in classifying samples of 4000 words (roughly the average size of an act). The possibility that combinations of these tests can be used successfully will be examined in Section 4.5.

4.3.1 Date of Composition and Style of Play

Variability within a writer's canon might be due to differences in the genre or date of play. If this were true the task of choosing a set of texts to characterize each author for comparison to *Henry VIII* and *The Two Noble Kinsmen* would become much more complex. A suitable statistical method for testing the hypothesis that a feature varies within subsets of samples is the *analysis of variance* (ANOVA). Assuming that the variances of the subsets are equal, this procedure estimates the overall variance by two methods. The ratio of these two estimates is distributed according to the variance-ratio distribution F, and a significant value refutes the null hypothesis that the subset means estimate the same population parameter. (Bailey provides a lucid description of the technique and its underlying assumptions [4].)

The effects of date and genre were tested only on the Shakespeare control set of 20 plays. Shakespeare's texts can be more accurately dated than Fletcher's, and the works fall more naturally into the categories of comedy, tragedy, history and romance. The exact dates of composition cannot of course be determined,

and even the exact order is disputed. To study the influence of date of composition the plays were divided into five groups as follows:

1. *The Comedy of Errors, Love's Labor's Lost, Two Gentlemen of Verona* and *The Taming of the Shrew*

2. *A Midsummer Night's Dream, Romeo and Juliet, King John* and *Richard II*

3. *The Merchant of Venice, 1 Henry IV, Much Ado About Nothing, Henry V* and *Julius Caesar*

4. *The Merry Wives of Windsor, Twelfth Night* and *All's Well That Ends Well*

5. *Macbeth, Coriolanus, Cymbeline* and *The Winter's Tale*

Table 2–1 on page 39 gives information on date and category. The divisions between groups might be considered somewhat arbitrary (especially between the first and second) but should be accurate enough to allow testing for chronological change.

The ANOVA tests were carried out on the proportion of occurrences in samples from the subsets listed. Again the act was used as the unit of observation. Tests for 10 of the 30 collocations and 2 of the 5 proportional pairs produced significant results at the 5% level. Table 4–7 lists these features with their F ratios in addition to the mean and standard error of the mean for each subset. Of the 10 features, the variation shown by 3 (*I have, to the* and *and the*) might show a linear increase according to period of composition. But for others no general pattern of development is discernible.

The same testing sequence was performed using an entire play as the unit of observation. In this case only 4 collocations yielded significant results; this reflects the decreased variance within groups for the larger samples. The results for *any+all* and *this+that* were still significant when rates from plays were used.

A similar procedure was employed to test for different rates of occurrence in comedies, tragedies, histories and romances. The division into subsets is accepted as follows:

Colloc. or Prop. Pair		Group Statistics				
F	prob.	$\overline{x}_1$	$\overline{x}_2$	$\overline{x}_3$	$\overline{x}_4$	$\overline{x}_5$
		$s_{\overline{x}_1}$	$s_{\overline{x}_2}$	$s_{\overline{x}_3}$	$s_{\overline{x}_4}$	$s_{\overline{x}_5}$
of the		7.61	7.13	8.65	8.45	15.02
10.36	5.12×10^{-7}	1.18	0.78	0.92	1.05	0.97
any+all		17.58	14.62	26.01	29.31	15.00
5.57	4.48×10^{-4}	2.93	2.70	2.32	3.91	2.19
I have		6.01	6.50	6.67	7.46	10.05
5.23	7.45×10^{-4}	0.63	0.62	0.59	0.86	0.87
of all		4.17	1.72	2.59	1.19	1.97
5.21	7.68×10^{-4}	0.79	0.39	0.35	0.29	0.39
this+that		31.82	42.56	35.92	36.24	34.78
4.61	1.90×10^{-3}	1.90	2.12	1.55	1.98	1.66
is a/an		8.78	5.29	8.43	11.08	8.37
4.39	2.65×10^{-3}	0.96	0.83	0.93	0.97	0.88
and X the		4.48	5.21	3.93	2.47	4.92
4.39	2.65×10^{-3}	0.53	0.43	0.47	0.43	0.42
to the		6.69	6.89	7.85	7.89	10.69
4.19	3.60×10^{-3}	0.64	0.62	0.83	0.84	0.89
is a		7.79	4.98	7.10	9.92	7.39
4.07	4.32×10^{-3}	0.85	0.79	0.83	0.76	0.73
the X X the		5.76	4.68	5.03	5.89	7.14
2.89	2.63×10^{-2}	0.66	0.54	0.48	0.71	0.43
the X and		6.70	5.27	6.75	6.40	4.98
2.86	2.75×10^{-2}	0.70	0.46	0.43	0.46	0.34
and the		2.62	2.50	2.82	4.18	3.86
2.48	4.90×10^{-2}	0.54	0.51	0.33	0.52	0.44

Means and their standard errors are calculated from proportions of occurrence (percentages) in acts of plays in each group. The first column of statistics corresponds to the earliest plays; the fifth column to the latest. The numbers of degrees of freedom for the F test are 4 and 96.

Table 4–7: Significant ANOVA results in Shakespeare by period of composition

Comedies: *The Comedy of Errors, Love's Labor's Lost, The Two Gentlemen of Verona, The Taming of the Shrew, A Midsummer Night's Dream, The Merchant of Venice, Much Ado About Nothing, The Merry Wives of Windsor, Twelfth Night* and *All's Well That Ends Well*

Tragedies: *Romeo and Juliet, Julius Caesar, Macbeth* and *Coriolanus*

Histories: *King John, Richard II, 1 Henry IV* and *Henry V*

Romances: *Cymbeline* and *The Winter's Tale*

Table 4–8 shows that 14 of 30 collocations show significant differences in means when acts are grouped according to these categories. As in grouping by period of composition, the number of significant results was much lower when rates and means were calculated from entire plays.

The rates for *of the* and *I have* are high for the group representing the latest plays and the romances, a correspondence which is hardly surprising considering that the romance plays were written last. In such cases it is impossible to determine from these tests which factor (date of composition or genre) is responsible for a high value in a particular act. Brainerd, studying pronoun rates and genre in Shakespeare's plays, used regression analysis and analysis of covariance procedures to interpret the interactions of date and genre [18]. Such sophisticated techniques were not employed in this study. The purpose of this analysis is to determine whether or not these features are affected by date or style.

An important qualification must be made in regard to the testing procedure just described. The ANOVA procedure assumes that the the variances for the groups are equal. This certainly may not be true for these divisions. However, the usual procedure for homogeneity of the variances, Bartlett's test, is sensitive to departures from normality [4]. Since the SAS goodness-of-fit tests showed that the rates for a number of these features are not distributed normally, this test was not carried out. Therefore it must be recognized that some of the results that appear significant may not be. However, for many of the variables the F ratios are very large and significance tests are unnecessary. The within-author inconsistency is evident from examination of the group means.

Collocation		Group Statistics			
F	prob.	$\overline{x}_1$	$\overline{x}_2$	$\overline{x}_3$	$\overline{x}_4$
		$s_{\overline{x}_1}$	$s_{\overline{x}_2}$	$s_{\overline{x}_3}$	$s_{\overline{x}_4}$
of the		8.17	11.50	7.71	14.51
6.91	2.89×10^{-4}	0.65	1.34	1.07	0.87
to a		2.58	1.20	0.95	1.04
6.04	8.16×10^{-4}	0.30	0.37	0.21	0.27
to a/an		2.75	1.32	1.16	1.27
5.65	1.31×10^{-3}	0.31	0.37	0.25	0.34
of X and		5.63	5.21	7.64	3.97
4.49	5.39×10^{-3}	0.47	0.43	0.56	0.66
I am		9.49	8.03	6.91	10.24
3.82	1.24×10^{-2}	0.51	0.64	0.69	0.93
a X and		5.01	5.82	7.20	4.36
3.43	2.01×10^{-2}	0.39	0.60	0.80	0.53
I have		6.46	8.21	7.29	9.64
3.25	2.51×10^{-2}	0.42	1.03	0.62	0.77
and all		1.14	0.91	1.84	0.69
3.25	2.51×10^{-2}	0.17	0.21	0.29	0.21
is a/an		9.25	7.60	5.93	9.24
3.21	2.64×10^{-2}	0.63	0.80	0.98	1.34
is a		8.17	6.73	5.29	8.00
3.11	2.99×10^{-2}	0.57	0.70	0.83	1.05
of a		3.80	2.16	4.08	2.63
2.99	3.47×10^{-2}	0.37	0.44	0.59	0.64
to be		5.66	4.46	4.93	7.58
2.73	4.81×10^{-2}	0.44	0.57	0.72	0.69
is the		8.20	6.29	7.65	4.62
2.72	4.87×10^{-2}	0.69	0.66	0.75	0.60

Means and their standard errors are calculated from proportion of occurrence (percentages) in acts of plays in each group. Groups 1–4 correspond to plays as follows: (1) Comedies, (2) Tragedies, (3) Histories and (4) Romances. The numbers of degrees of freedom for the F test are 3 and 97.

Table 4–8: Significant ANOVA results in Shakespeare by genre

The differences in the means shown in these tables cannot be attributed to a small variation in the counts due to a local anomaly or repetition. A small number of additional occurrences would probably most affect collocations or proportional pairs based on low-frequency counts. The following discussion examines the counts behind the extreme proportions of one such feature. While this detailed examination is not a formal analysis (and certainly not a complete one, for only one collocation is examined), it does indicate that small sections of texts are not responsible for the significant differences in the subset means shown in the tables.

The most infrequent feature in the two tables is *of all,* which has a proportion of 4.17% in Shakespeare's earliest plays compared to a figure of 1.19% in the last period. The high rate represents a total of 41 of 1121 keywords in the acts of the four plays for the earliest group. The low rate corresponds to 16 of 1300 keywords in the samples from the three plays making up Group 4. If the counts in these samples were proportional to the overall Shakespeare rate, the values in the two groups would be 24.2 and 28.1 occurrences respectively. The difference between the observed count of 41 and the expected 24.2 for the early group is quite large. Inspection of the occurrences of *of all* in the plays of this group reveals only one repetition: three occurrences of "of all the rest" occur in Act 1 of *The Two Gentlemen of Verona.* If this act is excluded from the analysis of variance, the mean for that group drops to 3.69 ($s = 0.67$) but the F ratio is still significant at 4.36. Thus the higher rate of *of all* is not due to just one or two samples but seems to be a characteristic of the entire subset. The means for the more frequent features should be less sensitive to the effects of a small change in the counts, and one can conclude that the ANOVA results presented do reveal meaningful variations due to date of composition and genre.

Each of the positional stylometry examinations of Shakespearean problems described in Section 3.3 assume that the features studied do not vary according to date or style of composition. None of the researchers adequately test this

assumption, accepting that the findings described in "To Couple Is the Custom" apply to their problem. The results presented in this section indicate that this cannot be safely assumed for every collocation and proportional pair. The ANOVA results indicate that Shakespeare's rate of use varies during his career for some features. Likewise, the differences between tragedies, comedies, histories and romances indicate that genre may affect the number of occurrences. This situation has not been observed in other stylometric studies (such as Morton and his colleagues' study of Scott and Fowles) and weakens the assertion that collocations and proportional pairs are generally consistent and subconscious habits of composition.

4.4 Correlation

The degree of variation within each writers' works inspires little confidence in the use of collocations and proportional pairs in an attribution study involving Shakespeare and Fletcher. While this is important for this study of authorship, another question surrounding these tests can also be examined. Any stylometric study that combines results from a number of tests must address the question of the statistical independence of tests. As described in Section 3.1.3 (page 97), Morton and his colleagues calculated correlation coefficients in chapters of one of Scott's novels, concluding correlation did not affect their procedures. O'Brien and Darnell do not report similar testing in their research. Merriam accepts that Morton has shown "that his stylometric tests are independent for the purposes required" [92, p. 176] although this has only been demonstrated for a single novel. No support for or against the independence of tests based on these features has been published since "To Couple Is the Custom."

There are two aspects of the concept of statistical independence. First, two events are not independent if there is a possibility that a single outcome counts as both of them. (In some cases the effects of this dependence may be quite small.) Second, two events are not independent if there is a significant tendency

for high values of one variable to be associated with high (or low) values of the other. The first restriction applies in cases such as the occurrence of "and all the" which would be counted as occurrences of both *and all* and *and* X *the.* In addition, if 2×2 contingency tables are being evaluated using χ^2 or exact tests, then each test is based both on the number of keywords characterized by the feature and by those occurrences that are not. In this case tests based on the same keywords will not be independent. To get around this problem an $n \times 2$ table can be used, or the test can be redefined. (For example, Merriam counts the number of occurrences of *I* followed by *am* in one 2×2 table. Then in counting *I have* he subtracts the keyword count for *I am* from the total count of *I* to obtain a modified contingency table for *I have* [92, p. 172]).

The second type of independence can be measured empirically using statistical tests of correlation. The analysis of correlation in a set of values is a particularly laborious process because the statistic should be calculated for every possible pair of variables. For thirty variables there are exactly $(30 \times (30-1))/2 = 435$ combinations: a prohibitive amount of calculation without a computer package like SAS. The Pearson product-moment coefficient is the most commonly used statistic of correlation (and was used by Morton in his analysis of collocations in Scott). The significance of this statistic depends on the number of observations used to calculate it; in addition, most published tables of these significance levels are based on the assumption that the two variables are approximately bivariate normal.

As noted in Section 4.3 many collocation rates are not normally distributed. Two common statistics of rank correlation are available when normality cannot be assumed or when one (or both) of the variables being tested is simply a ranking. Kenny [61] describes the Spearman rank correlation coefficient, while Bailey discusses Kendall's τ, noting that the latter has certain practical and theoretical advantages [4]. Significance levels for the Spearman coefficient are computed as for Pearson's ρ (as described in Snedecor and Cochran [151, p. 185]). Bailey provides a small table indicating significance levels for Kendall's τ for

small numbers of observations. More complete tables can be found in other recent compilations of tables, such as Powell's [121].

SAS was used to calculate all three correlation coefficients for the proportion of occurrence of each feature in the acts of both Fletcher and Shakespeare. The compiled results for Kendall's τ are presented for both authors in Tables 4–9 and 4–10 (starting on page 173). Each collocation and proportional pair is followed by a list of the other variables with which it significantly correlates at the 5% level of significance. If a list item is printed in *italics* then the correlation is negative. The tables show every instance of significant correlation, even for variants (such as *to a* and *to a/an*) where the counts are often almost identical. In some cases a particular test is just significantly associated with one member of a variant pair and not the other. For example, in Fletcher *I have* significantly correlates with *it is* but not with *it is/was*.

To judge the extent of significant combinations, one can disregard one member of the variant pairs and count the proportion of total combinations that are significant. Choosing the four "followed by" *a/an* variants and *it is/was* over *it is*, the remaining 25 collocation and 5 proportional pair tests can be combined in 435 different ways. For Fletcher 29 pairs (6.7% of the total) are significantly correlated at the 5% level; for Shakespeare the number is larger, 41 pairs (9.4%). Results for the Spearman coefficient are almost identical to τ results presented. Although bivariate normality cannot safely be assumed, the Pearson coefficients were also tabulated; these result in 16 more significant pairs in Shakespeare and 8 more in Fletcher.

No pattern emerges when the lists for both writers are compared. Indeed, they only share two significantly correlated pairs: *in the* with *of the*, and *and the* with *without+with*. Significant statistical correlation does not necessarily imply cause and effect relationships among correlated variables, although one could propose reasonable-sounding theories for the correlation of *I did* and *I do*, or for *by the*, *of the* and *in the* in Shakespeare.

The important point is that a number of collocation and proportional pair

tests are correlated in both Shakespeare and Fletcher. If one were using any correlated pairs to test for consistency within a single writer's works, one *could not* validly combine significance test probabilities as if they were independent. This certainly applies to the method of multiplying the probabilities resulting from a sequence of χ^2 or exact tests, as practiced by Metz and Morton [94] and Merriam [92]. Clearly multivariate statistical techniques that allow for association among variables should be employed. Such techniques are the subject of Chapter 6.

For each collocation and proportional pair, the following list indicates the other features in Fletcher with which it is correlated at the 5% level of significance. The tests are based on proportion of occurrence counted in 30 acts. The rank correlation coefficient used is Kendall's τ. List items printed in italics indicate a negative correlation.

and all:	—
and the:	the X and, of a, of a/an, without+with
and X the:	a X and, *an+a*, *this+that*
a X and:	in a/an, and X the, of the
a X of:	I do
by the:	I am, *I did*, of a, of a/an
I am:	by the, without+with
I did:	in the, *by the*, *to a*, *is a*, *is a/an*
I do:	to be, a X of
I have:	to the, *it is*
in a:	in a/an
in a/an:	in a, a X and
in the:	of the, I did, the X X the
is a:	is a/an, the X and, *I did*, *this+that*
is a/an:	is a, the X and, *I did*, *this+that*
is the:	without+with
it is:	it is/was, *of the*, *I have*, *to the*, no+not
it is/was:	it is, *of the*
of a:	of a/an, and the, by the
of a/an:	and the, by the, of a
of all:	to be
of the:	in the, *it is/was*, *it is*, a X and
of X and:	—
the X and:	is a/an, and the, is a, an+a, *this+that*
the X the:	—
the X X the:	in the, an+a
to a:	to a/an, *I did*
to a/an:	to a
to be:	I do, of all
to the:	I have, it is
an+a:	the X and, the X X the, *and* X *the*
any+all:	*no+not*
no+not:	*any+all*, it is
this+that:	*is a*, *is a/an*, *and* X *the*, *the* X *and*
without+with:	I am, and the, is the

Table 4–9: Correlated collocations and proportional pairs in Fletcher

For each collocation and proportional pair, the following list indicates the other features in Shakespeare with which it is correlated at the 5% level of significance. The tests are based on proportion of occurrence counted in 101 acts. The rank correlation coefficient used is Kendall's τ. List items printed in italics indicate a negative correlation.

and all:	*the* X X *the*, *of the*, *in a/an*, *in a*, *is a/an*, *is a*, *to be*, and X the
and the:	the X X the, the X the, in the, to the, without+with
and X the:	*it is*, in the, and all, *it is/was*, *to a*
a X and:	this+that
a X of:	this+that
by the:	of the, in the, of a
I am:	*of* X *and*
I did:	I do, *the* X X *the*
I do:	I did
I have:	of the
in a:	in a/an, *and all*, to a, is the, *in the*, to a/an
in a/an:	in a, *and all*, is the, *in the*, to a
in the:	to the, and the, the X the, of the, by the, and X the, of a, *in a/an*, is a/an, *in a*, without+with
is a:	is a/an, the X X the, *and all*, *an+a*, any+all
is a/an:	is a, the X X the, *and all*, in the, any+all
is the:	in a/an, in a, any+all
it is:	it is/was, *and* X *the*, *this+that*
it is/was:	it is, *and* X *the*
of a:	of a/an, in the, to be, by the
of a/an:	of a
of all:	—
of the:	the X X the, the X the, *and all*, in the, to the, by the, *the* X *and*, I have
of X and:	the X and, *I am*, *the* X X *the*
the X and:	of X and, *of the*
the X the:	and the, in the, of the, the X X the, to the
the X X the:	of the, and the, is a/an, *and all*, is a, the X the, *I did*, *of* X *and*
to a:	to a/an, in a, in a/an, *and* X *the*
to a/an:	to a, in a, without+with
to be:	*and all*, of a, without+with
to the:	in the, of the, the X the, and the
an+a:	*is a*
any+all:	is a/an, is a, is the
no+not:	—
this+that:	a X and, *it is*, a X of
without+with:	to be, and the, in the, to a/an

Table 4–10: Correlated collocations and proportional pairs in Shakespeare

4.5 Application to the Test Set

Clearly collocations and proportional pairs in these texts are not as well-suited for Shakespearean authorship studies as some previous studies have indicated. The large within-writer variances would render any test very unreliable when used on its own. The combination of information from a number of tests, however, might produce the right result. After all, these tests have been used in several studies in which the method has been validated on some control samples. Have the scholars behind these studies been entirely misled? This section will evaluate the observed misclassification rate of a study based on the tests examined in this chapter.

In interpreting the results of significance tests, the authors of "To Couple Is the Custom" adopt the relative frequency interpretation of probability. Recognizing the possibility of a Type 1 error, they outline the rationale for combining the results of significance tests:

> In a set of of twenty samples one difference significant at the 5% level will be expected to occur. However, the best defense against being misled by this type of error is to use a number of independent tests. Two such tests will combine to mislead, at the 5% level, only once in twenty times twenty trials, i.e. once in 400 trials. Half a dozen independent tests will combine to mislead only once in several million trials [102, p. 12].

Such reasoning (with the relative frequency interpretation of probability behind it) leads to the multiplication of probabilities from χ^2 test tests. One way of evaluating the effectiveness of such methods of combining tests is to apply the procedure to a number of test samples for which the authorship is known.

The six plays of the test set will be used to estimate the misclassification rate of the method of multiplying significance test results used in Metz and Morton's study of *Titus Andronicus* [94] and Merriam's study of the Huntingdon plays [92]. The first step is to choose a set of collocations and proportional pairs,

based on the analysis of differences between the two writers compared to the within-author variation. However the graphs of the standard deviations and interquartile ranges indicate that occurrences in each author's texts generally overlap the other's, even when the t test (which takes standard deviation into account) indicates a significant difference in mean. Although none are as consistent as one would have hoped, a set of variables with highly significant values of t' was chosen from the top of Table 4–4. These features are: *I did, no+not, of all, to the, and all* and *a* X *of.*

Since 2×2 tables are being evaluated, only one collocation for a given keyword was chosen, thus avoiding the first sort of statistical dependence described above. By chance, none of these six variables are significantly correlated, so the results of the individual significant tests are independent. For each test a 2×2 contingency was formed and the probability from Fisher's exact test computed. For each author the probabilities for each significance test were multiplied, and the products were divided to form a likelihood ratio. This procedure was first performed on large samples: the six complete plays in the test set.

All six plays were correctly classified using the above method with this set of variables. Several of the likelihood ratios are truly astronomical, but the ratio for *As You Like It* is only 2.7 to 1 in favor of Shakespeare. Next, the same set of tests was applied to the individual acts. Table 4–11 shows the logs of the products of the probabilities and the likelihood ratios for each sample. Two acts of *As You Like It* and two acts of *Valentinian* are misclassified. This suggests an error rate of about 13% for these 30 samples, which have an average size of over 4000 words. Likelihood ratios for 7 of the correctly classified acts of Shakespeare are less than 10 to 1, and 3 of these are less than 3 to 1. These results further suggest that this procedure and the features tested do not always provide absolute discrimination with a large margin for error. (Other combinations of the best markers in Table 4–4 were tested in the same way. The attributions were sometimes slightly different but the number of misclassifications and borderline cases was comparable.)

Sample	$\sum$ ln *prob* Fletcher	Shakespeare	Likelihood ratio	
Ant	-23.02120	-5.09550	$6.096\times10^{+07}$	Sh
1	-7.83940	-5.42230	$1.121\times10^{+01}$	Sh
2	-11.14660	-4.73210	$6.106\times10^{+02}$	Sh
3	-10.36510	-5.00150	$2.135\times10^{+02}$	Sh
4	-10.34910	-8.40670	6.975	Sh
5	-5.60690	-1.28260	$7.551\times10^{+01}$	Sh
AYL	-13.82520	-12.82290	2.725	Sh
1	-7.10150	-4.88250	9.198	Sh
2	-10.76760	-8.96790	6.048	Sh
3	-4.66000	-6.78370	8.362	Fl
4	-6.84010	-4.49630	$1.042\times10^{+01}$	Sh
5	-4.87920	-6.99740	8.316	Fl
R3	-30.72690	-7.34880	$1.422\times10^{+10}$	Sh
1	-16.29920	-6.65550	$1.542\times10^{+04}$	Sh
2	-4.86640	-3.61850	3.483	Sh
3	-10.26890	-3.96740	$5.454\times10^{+02}$	Sh
4	-13.78630	-5.96100	$2.503\times10^{+03}$	Sh
5	-12.88440	-8.69080	$6.626\times10^{+01}$	Sh
Tem	-18.68420	-6.15030	$2.776\times10^{+05}$	Sh
1	-8.17650	-7.80090	1.456	Sh
2	-5.12070	-4.99210	1.137	Sh
3	-7.62500	-1.82600	$3.300\times10^{+02}$	Sh
4	-2.69980	-2.43090	1.309	Sh
5	-9.72890	-2.92250	$9.036\times10^{+02}$	Sh
Thom	-7.10850	-42.66800	$2.775\times10^{+15}$	Fl
1	-3.98880	-14.32670	$3.088\times10^{+04}$	Fl
2	-4.99620	-13.08470	$3.257\times10^{+03}$	Fl
3	-4.80930	-13.11630	$4.052\times10^{+03}$	Fl
4	-7.36570	-11.33160	$5.277\times10^{+01}$	Fl
5	-6.83850	-13.64890	$9.072\times10^{+02}$	Fl
Vale	-9.66380	-18.48660	$6.787\times10^{+03}$	Fl
1	-6.39860	-15.03710	$5.645\times10^{+03}$	Fl
2	-5.23910	-3.28840	7.034	Sh
3	-7.80950	-5.85430	7.065	Sh
4	-6.66410	-9.61100	$1.905\times10^{+01}$	Fl
5	-4.40160	-9.13590	$1.138\times10^{+02}$	Fl

For each writer, a probability was calculated (using Fisher's exact test) for *I did, no+not, of all, to the, and all* and *a* X *of*. For reasons of accuracy the probabilities were converted to natural logarithms in order to calculate a likelihood ratio for each classification. The results for each play are followed by those for its individual acts.

Table 4–11: Probability sums for six tests on test-set samples

Several of the features used in this analysis do not meet the recommended minimum requirement of 5 expected occurrences, even in samples averaging 4000 words. Statistical considerations dictate the use of Fisher's exact test over χ^2 in this situation, but results based on infrequent features may be affected by local anomalies. The procedure was repeated using the first four variables in Table 4–4 that meet this minimum requirement: *to the, a* X *of, no+not* and *this+that.* Three plays and 11 of the acts were misclassified using these features. This does not compare favorably with the 13% misclassification rate for acts achieved with the less frequent tests.

Merriam uses this testing procedure with infrequent features on small samples in his study of the Huntingdon plays. He relies on the redundancy of information in a series of tests to compensate for the increased variability in small samples, and uses Fisher's exact test to avoid the problems of χ^2 tests with small numbers. The variables used in Table 4–11 were counted in individual scenes of the test-set plays to see how the error rate increased with a reduction of sample size. Only 16 scenes in the test set are larger than 2000 words; 5 of these are misclassified by this procedure. In decreasing the minimum size to 1500 words the rate of misclassification increases to 32% (9 of 28). If scenes of 1000 or more words are tested, 14 of 50 are attributed incorrectly, a rate of 28%.

Describing this figure as an "error" rate is misleading; many of the likelihood ratios are very close to 1.0. In their studies Morton, Merriam and most others emphasize that an attribution will only be made if the weight of evidence is strong. These results demonstrate that the method and the variables behind it are not as powerful as previous studies have concluded. This demonstration can be criticized, since no attempt has been made to trace the cause of misclassifications to local anomalies or repetitions in the texts. But if a sample the size of an act is small enough to be affected by such "less-random" word usages, then these tests and methods are not very valuable for the analysis of Jacobean dramatic questions.

4.6 Conclusions

Many researchers have focused attention on collocations, in part because of the assertion by Morton and his colleagues that occurrences of collocations vary "much less within a writer and much more between writers than the occurrence of either word which makes up the collocation" [102, p. 17]. A careful reader will have noticed that the collocations and proportional pairs that scored best as discriminators (according to the t tests) are composed of a small number of function words: *no, did, all, of, to, the, a* and *and.* When t tests are applied to the rates for these words in acts, this supposed general characteristic of collocations is not evident. Table 4–12 on page 182 presents graphs showing these t test results, the standard deviations and interquartile ranges. Inspection shows that in most cases the words alone discriminate between acts of Shakespeare and Fletcher at least as well as the collocations do; even better if one judges by the probabilities associated with these t values. If words like *a, and, of, the* and *to* could be used as discriminators, shorter samples could be tested because of high rate of occurrence of these function words.

At this stage I have demonstrated that the methods developed around these collocations and proportional pairs are not especially useful in an authorship study involving Shakespeare and Fletcher. The variation within the plays of the two dramatists overwhelms the differences between their works in most cases. In using the *best* of these tests a misclassification rate of 13% was obtained when the 30 acts in the control set were tested. More complex procedures based on these variables (such as the higher-dimension contingency tables suggested by Smith and O'Brien and Darnell) were not evaluated. The graphs showing the internal variation within acts of Shakespeare and Fletcher (page 160) are discouraging. Moreover, the statistics for the function words in Table 4–12 suggest that one might be more successful using these as variables.

Positional tests based on words labeled according to grammatical class might

be more successful than simple collocation tests. The "followed by adjectives" collocations used by Morton and his associates in some studies are a step towards a positional stylometry of grammatical word classes. At the moment, however, tagging is a manual process for the most part and is impractical for the large number of samples of text required to validate authorship methods. Software "tagging" systems are developing rapidly (for example, the CLAWS system [79] mentioned at the end of Chapter 3). When such systems have been adapted for use with Early Modern English, study of the frequency and pattern of occurrence of grammatical classes may result in many useful tests of authorship.

Another possible approach would be to examine the positions of words in blocks delimited by common words. Such a procedure has been described by Michaelson and Morton in "The Spaces in Between" [100]. This method involves the comparison of the frequency distributions for two authors after determining a model for the underlying distribution of word occurrences. This technique could be applied to the current problem; however, simple word rates show promise as discriminators (as shown by Figure 4–12). These variables are more easily compared than frequency distributions; in addition, distribution-free techniques can be employed in their analysis. This is certainly an advantage in view of the problems surrounding the description of distributions of literary features (discussed in Section 4.3).

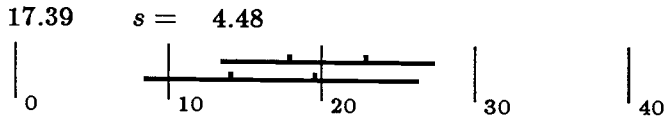
The reader should note that the negative conclusions regarding the collocations and proportional pairs examined in this chapter should not be taken as a evidence that they will not work when applied to other writers in English. One should not ignore the positive results in some other studies. However, the Shakespeare samples used in this study represent the largest body of text written by one author in which a set of collocations and proportional pairs has been analyzed. Of course, the observed internal variation within these samples may be due to the nature of Jacobean dramatic texts or to the characteristics of Early Modern English. Clearly more complete sampling of several writers' works is

required to re-evaluate the usefulness of collocation and proportional pair tests for general use.

Note: The bars on the right indicate $\overline{x} \pm 2s$. The ticks on top of the bars mark the 1st and 3rd quartiles. Word rates are expressed in units of 1000 words.

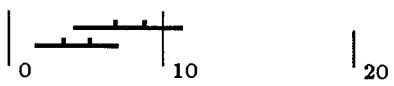
a $t' = 3.92 \quad \nu' = 59.70 \quad p = 2.33 \times 10^{-4}$

Fl: $\overline{x} = 20.43 \quad s = 3.48$
Sh: $\overline{x} = 17.39 \quad s = 4.48$



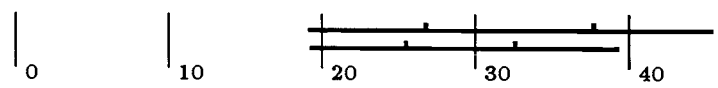
all $t' = 9.57 \quad \nu' = 39.35 \quad p = 8.75 \times 10^{-12}$

Fl: $\overline{x} = 7.73 \quad s = 1.76$
Sh: $\overline{x} = 4.38 \quad s = 1.35$



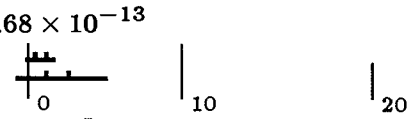
and $t' = 2.29 \quad \nu' = 39.33 \quad p = 0.0275$

Fl: $\overline{x} = 32.25 \quad s = 6.57$
Sh: $\overline{x} = 29.27 \quad s = 5.02$



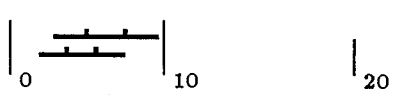
did $t' = -8.02 \quad \nu' = 127.69 \quad p = 8.68 \times 10^{-13}$

Fl: $\overline{x} = 0.80 \quad s = 0.48$
Sh: $\overline{x} = 2.18 \quad s = 1.49$



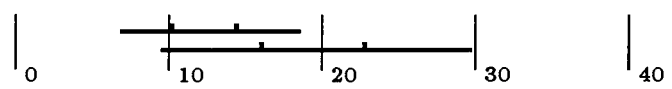
no $t' = 4.65 \quad \nu' = 41.00 \quad p = 3.43 \times 10^{-5}$

Fl: $\overline{x} = 6.25 \quad s = 1.71$
Sh: $\overline{x} = 4.66 \quad s = 1.40$



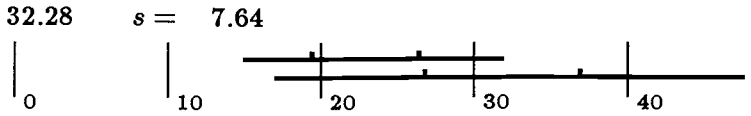
of $t' = -9.36 \quad \nu' = 82.80 \quad p = 1.68 \times 10^{-14}$

Fl: $\overline{x} = 12.73 \quad s = 2.94$
Sh: $\overline{x} = 19.62 \quad s = 5.07$



the $t' = -8.15 \quad \nu' = 86.96 \quad p = 2.63 \times 10^{-12}$

Fl: $\overline{x} = 23.45 \quad s = 4.24$
Sh: $\overline{x} = 32.28 \quad s = 7.64$



to $t' = -5.17 \quad \nu' = 66.66 \quad p = 2.36 \times 10^{-6}$

Fl: $\overline{x} = 19.33 \quad s = 2.51$
Sh: $\overline{x} = 22.34 \quad s = 3.59$

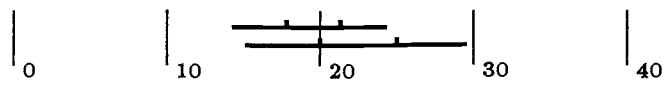


Table 4–12: Internal variation of some word rates measured in acts

Chapter 5

Finding Common Words that Discriminate

As noted in the concluding section of the last chapter, some function words appear to discriminate between Shakespeare and Fletcher more effectively than the other variables examined thus far. To determine if word-rate variables can be used to analyze *Henry VIII* and *The Two Noble Kinsmen,* one must first discover which words best meet the criteria for habits of authorship (discussed in Chapter 1 and Section 3.1.2). This chapter describes the procedures used in discovering these *marker words.* In addition to individual words, some common grammatical word classes (for example, pronouns) are examined as a group to determine if the two playwrights use such groups at different rates.

5.1 Individual Word Rates

Relative frequencies for words are often measured in terms of rates per thousand, and this convention has been adopted in this study. Measuring word occurrence by rates is convenient since it allows one to compare data from samples of different length (measured in number of words). However, using rates instead of counts does raise difficulties at times. The fact that a 500 word sample contains zero occurrences of a particular word might be much less significant than the

fact that the word is absent from a 2000 word sample, although the word-rate measurement for both is identical. Although word rates appear to be continuous variables, they are in fact discrete for a sample of a given length. For example, in samples of 500 words the variable "rate of *dare*" can only take on values of 0.0, 2.0, 4.0.... Therefore the rate in such a sample could never equal the overall rate for *dare* in Fletcher, which is 1.27 per thousand.

Some forms of statistical analysis (including the discriminant analysis procedures introduced later in this study) may attach great significance to a word-rate value in a small sample when in fact this might be due to the insertion or deletion of a single occurrence of the word. On the other hand, statistical procedures that allow for non-normal distributions and correlated variables cannot easily deal with counts in samples of different lengths. This problem was frequently encountered by Mosteller and Wallace [113]. While their main study is based on word counts and takes paper length into account, a number of their secondary analyses were limited to papers containing about the same number of words. Most authorship problems (including the one under consideration in this study) require that samples of different length be compared. In the sections that follow, word rates will be treated as if they were continuous variables. However, results for small samples will be examined carefully in an attempt to determine if small changes in the word counts could seriously affect any conclusions.

5.1.1 Distinctiveness Ratios

Ellegård introduced the idea of a word's *distinctiveness ratio,* which is the ratio of the relative frequency in one writer's works to the relative frequency in another writer's works [31]. To illustrate the distinctiveness ratio, consider the word *too*. In the six Fletcher plays *too* occurs 604 times, a rate of 4.52 per thousand words. In the 20 Shakespeare plays the count is 707 occurrences, a rate of 1.66 per thousand words. The distinctiveness ratio for Fletcher is $4.52/1.66 = 2.72$. For a given author, words used at different rates are often referred to as "plus" or

"minus" words according to whether they are used more often by one author in comparison to another.

The computer's power is crucial in discovering words that may be good markers of authorship. Having produced word counts for each play in the control set, the machine was used to combine these into an overall word list for each dramatist. These two lists were then merged; the counts and rates for each word used by both writers were listed on a single line. Words used on average less than five times in a play by either author were removed from this list and processed at a later stage. The distinctiveness ratios were calculated so that every word was marked as a "plus" word for one playwright or the other, and the file was sorted according to these values. The beginning of this file, listing the words with the highest ratios, is given in Table 5–1.

The three words in the list with the highest distinctiveness ratios are *ye, hath* and *them.* These three words and *'em* (which has a distinctiveness ratio of 41.4 for Fletcher but is not listed in Table 5–1 because Shakespeare uses it less than 5 times per thousand words) have been noted in the past by scholars and have figured prominently in previous studies of Fletcher authorship questions. (The relevant studies will be reviewed in Chapter 7.)

Some words that appeared in the earliest versions of this list reflected a difference in spelling rather than rate of use. For each function word with a distinctiveness ratio larger than 1.25, the complete word lists for both authors were examined to see if spelling variants were responsible for the apparent difference. Where this was the case the word was eliminated from consideration. For example, *again* appeared to be a function word with a high distinctiveness ratio, but this difference disappeared when its counts were merged with its variants (such as *againe* and *agen*). The variants for the most common function words should have been standardized at an earlier stage by the program REPLACE. Variants that were discovered at this stage were added to the replacement lists, and the word counts were generated again.

	Fletcher		Shakespeare			
	Count	Rate	Count	Rate	Dist.	Ratio
ye	1911	14.35	101	0.24	60.6	Fl+
hath	32	0.24	1036	2.43	10.1	Sh+
them	41	0.31	1032	2.42	7.9	Sh+
honour	178	1.34	110	0.26	5.2	Fl+
dare	170	1.28	107	0.25	5.1	Fl+
has	297	2.23	188	0.44	5.1	Fl+
sure	216	1.62	154	0.36	4.5	Fl+
gentlemen	138	1.04	107	0.25	4.1	Fl+
lord	80	0.60	945	2.22	3.7	Sh+
being	32	0.24	357	0.84	3.5	Sh+
woman	187	1.40	175	0.41	3.4	Fl+
honest	132	0.99	135	0.32	3.1	Fl+
aye	42	0.31	388	0.91	2.9	Sh+
too	600	4.51	707	1.66	2.7	Fl+
did	107	0.80	893	2.09	2.6	Sh+
beleeue	92	0.69	114	0.27	2.6	Fl+
speake	69	0.52	543	1.27	2.5	Sh+
lost	88	0.66	115	0.27	2.4	Fl+
himself	31	0.23	242	0.57	2.4	Sh+
found	78	0.59	104	0.24	2.4	Fl+
keepe	32	0.24	245	0.57	2.4	Sh+
king	72	0.54	550	1.29	2.4	Sh+
downe	38	0.28	285	0.67	2.3	Sh+
still	182	1.37	249	0.58	2.3	Fl+
which	165	1.24	1230	2.89	2.3	Sh+

Table 5–1: Words with large distinctiveness ratios in control texts

5.1.2 *t* Tests

Distinctiveness ratios are useful for finding relatively infrequent words that are used at very different rates. However, the difference between the two authors' rates for a common word can be large but still result in a low distinctiveness ratio. For example, the rates for *the* in the Shakespeare and Fletcher controls are 32.6 and 23.7 respectively. The difference is almost 9 words per thousand but the distinctiveness ratio is only 1.38 for this Shakespeare "plus" word. While the ratio of the relative frequencies is a useful measure for isolating possible markers of authorship, it takes no account of the difference between rates but only of their relative magnitudes. Even the difference between rates is not entirely satisfactory as a measure of usefulness, since it does not take into account the within-author variation.

To determine if two means are significantly different, statisticians use t tests to measure the difference between the observed means in terms of the overall variance. This test was used in the preceding chapter to determine which collocations and proportional pairs best discriminate between samples of Shakespeare and Fletcher. More information than the overall rate is required in order to use this statistic to recognize possible markers of authorship. The word rates must be measured in a number of samples in order to obtain a mean and a sample variance for each author. Ideally one would calculate these statistics from text samples that are about the same length as the disputed scenes. Making counts and calculating the statistics for every word in each act or scene in the control set would require a great deal of computer resources. Although this might not have been impossible, the existing word-count lists for individual plays were used in this first stage of the selection process. Possible markers found at this stage were later counted in smaller samples.

The overall rates for Shakespeare and Fletcher were averaged. For the 500 most frequent words the statistic t' (Equation 4.1 on page 152) was calculated from the means and variances of the 20 Shakespeare and 6 Fletcher control set plays. If a word could be considered a function word and had a value of t'

greater than 2.0, it was selected for further examination. Calculating a value for the Fletcher sample variance from only 6 observations is not very good statistical practice. It is unlikely that this will have any great effect in this initial screening.

5.1.3 Frequency Distributions of Word Rates in Scenes

Next, potential marker words were counted in acts and scenes; the t statistics calculated from these samples were usually larger than the values calculated from the rates counted by play. The probability associated with t' can be used to assess the magnitude of the difference in average word rates in the two writers. However, if the number of degrees of freedom associated with the t' values is large, it is statistically acceptable to compare the values directly. The value of t' is approximately a standard normal deviate when the number of degrees of freedom is greater than about 60. Recall that the value for the degrees of freedom for t' is calculated on the number of samples and the standard deviation for each author according to (4.2) on page 152. In most of the word-rate comparisons made in this study, these values were large enough that probabilities were often not calculated.

For many of these words frequency distributions for the rates in scenes were tabulated. Distributions for *a, in, of, that, the* and *too* are shown in Table 5–2 (which begins on page 190). Each word rate was measured in scenes containing at least 1000 words, and the numbers of scenes with rates falling in fixed intervals are indicated. The counts are also shown as percentages of the total number of scenes for each author (168 for Shakespeare and 54 for Fletcher). Cumulative percentages indicate what proportion of the scenes have rates below the upper boundary of any interval.

Inspection of a number of these frequency distributions showed that the distributions for many words with a significant t test result overlap to a large extent. This is not really surprising; the t test is used to determine if two means are significantly different, but one imagines that authorship studies need variables that differ a great deal. For example, the average rates for *that* are significantly

different in the two authors' samples; the probability associated with $t' = -2.01$ and 84 degrees of freedom is 0.048. Yet the two rates do not appear to be all that different (13.6 for Shakespeare and 12.4 for Fletcher), and the authors' frequency distributions overlap considerably. On the other hand, the two words with the largest t' values, *in* and *of,* show more promise. For *in,* over 90% of the Fletcher scenes have rates lower than 12; about 57% of the Shakespeare scenes have higher rates. Less than 4% of the Fletcher scenes have a rate of *of* higher than 20, while almost half of Shakespeare's scenes do.

These frequency distributions show that even the common words with very large differences in their rate of occurrence are not effective discriminators on their own. Used in combination they could yet prove useful. Mosteller and Wallace found that the high-frequency function words performed better than any other set of marker words they analyzed in *The Federalist* papers. But at this stage of the selection process it seems clear that words that do not have very large t' values could be ruled out. However, a number of these words with lower values were retained, solely because they are frequent: *a, and, as* and *to,* for example. Later testing confirmed the usefulness of the t statistic; most of these did not prove useful discriminators and were eliminated in later stages of the analysis. Table 5–3 lists the set of 23 marker words retained, with the mean rate and standard error for each author listed with the t statistics.

Frequency Distribution for *a:*

Rate per 1000 words	Shakespeare Num.	%	Cum. %	Fletcher Num.	%	Cum. %
4–8	3	1.8	1.8	0	0.0	0.0
8–12	27	16.1	17.9	3	5.6	5.6
12–16	41	24.4	42.3	8	14.8	20.4
16–20	47	28.0	70.2	13	24.1	44.4
20–24	22	13.1	83.3	14	25.9	70.4
24–28	16	9.5	92.9	12	22.2	92.6
28–32	8	4.8	97.6	4	7.4	100.0
32–36	2	1.2	98.8	0	0.0	100.0
36–40	2	1.2	100.0	0	0.0	100.0
	Mean:		17.8	Mean:		20.5
	Std Dev.:		6.16	Std Dev.:		4.96

$t' = 3.31$, df $= 109.2$, prob. $= 1.27 \times 10^{-3}$

Frequency Distribution for *in:*

Rate per 1000 words	Shakespeare Num.	%	Cum. %	Fletcher Num.	%	Cum. %
2–4	1	0.6	0.6	0	0.0	0.0
4–6	3	1.8	2.4	10	18.5	18.5
6–8	6	3.6	6.0	16	29.6	48.1
8–10	24	14.3	20.2	16	29.6	77.8
10–12	38	22.6	42.9	7	13.0	90.7
12–14	28	16.7	59.5	2	3.7	94.4
14–16	28	16.7	76.2	1	1.9	96.3
16–18	21	12.5	88.7	0	0.0	96.3
18–20	6	3.6	92.3	2	3.7	100.0
20–22	9	5.4	97.6	0	0.0	100.0
22–24	2	1.2	98.8	0	0.0	100.0
24–26	2	1.2	100.0	0	0.0	100.0
	Mean:		13.4	Mean:		8.5
	Std Dev.:		4.06	Std Dev.:		2.99

$t' = -9.43$, df $= 120.0$, prob. $= 4.88 \times 10^{-15}$

Table 5–2: Frequency distributions for *a* and *in* in scenes of at least 1000 words

Frequency Distribution for *of:*

Rate per 1000 words	Shakespeare Num.	%	Cum. %	Fletcher Num.	%	Cum. %
0–4	1	0.6	0.6	0	0.0	0.0
4–8	5	3.0	3.6	7	13.0	13.0
8–12	14	8.3	11.9	17	31.5	44.4
12–16	33	19.6	31.5	18	33.3	77.8
16–20	34	20.2	51.8	10	18.5	96.3
20–24	39	23.2	75.0	2	3.7	100.0
24–28	24	14.3	89.3	0	0.0	100.0
28–32	10	6.0	95.2	0	0.0	100.0
32–36	5	3.0	98.2	0	0.0	100.0
36–40	2	1.2	99.4	0	0.0	100.0
40–44	1	0.6	100.0	0	0.0	100.0

Mean: 19.8 Mean: 12.7
Std Dev.: 6.71 Std Dev.: 3.92
$t' = -9.53$, df $= 155.1$, prob. $= 0.0$

Frequency Distribution for *that:*

Rate per 1000 words	Shakespeare Num.	%	Cum. %	Fletcher Num.	%	Cum. %
3–6	2	1.2	1.2	0	0.0	0.0
6–9	10	6.0	7.1	12	22.2	22.2
9–12	47	28.0	35.1	15	27.8	50.0
12–15	55	32.7	67.9	11	20.4	70.4
15–18	35	20.8	88.7	10	18.5	88.9
18–21	15	8.9	97.6	5	9.3	98.1
21–24	3	1.8	99.4	1	1.9	100.0
24–27	1	0.6	100.0	0	0.0	100.0

Mean: 13.6 Mean: 12.4
Std Dev.: 3.53 Std Dev.: 3.80
$t' = -2.01$, df $= 84$, prob. $= 4.76 \times 10^{-2}$

Table 5–2 (*cont.*)**:** Frequency distributions for *of* and *that* in scenes of at least 1000 words

Frequency Distribution for *the:*

Rate per 1000 words	Shakespeare Num.	%	Cum. %	Fletcher Num.	%	Cum. %
5–10	0	0.0	0.0	2	3.7	3.7
10–15	2	1.2	1.2	6	11.1	14.8
15–20	11	6.5	7.7	10	18.5	33.3
20–25	25	14.9	22.6	15	27.8	61.1
25–30	37	22.0	44.6	15	27.8	88.9
30–35	40	23.8	68.5	4	7.4	96.3
35–40	18	10.7	79.2	1	1.9	98.1
40–45	17	10.1	89.3	1	1.9	100.0
45–50	11	6.5	95.8	0	0.0	100.0
50–55	3	1.8	97.6	0	0.0	100.0
55–60	3	1.8	99.4	0	0.0	100.0
60–65	1	0.6	100.0	0	0.0	100.0

Mean: 32.3 Mean: 22.8
Std Dev.: 9.32 Std Dev.: 6.38
$t' = -8.42$, df $= 130.2$, prob. $= 9.37 \times 10^{-14}$

Frequency Distribution for *too:*

Rate per 1000 words	Shakespeare Num.	%	Cum. %	Fletcher Num.	%	Cum. %
0	19	11.3	11.3	1	1.9	1.9
0–2	91	54.2	65.5	6	11.1	13.0
2–4	45	26.8	92.3	20	37.0	50.0
4–6	12	7.1	99.4	13	24.1	74.1
6–8	1	0.6	100.0	10	18.5	92.6
8–10	0	0.0	100.0	3	5.6	98.1
10–12	0	0.0	100.0	0	0.0	98.1
12–14	0	0.0	100.0	1	1.9	100.0

Mean: 1.72 Mean: 4.52
Std Dev.: 1.37 Std Dev.: 2.49
$t' = 7.90$, df $= 63.4$, prob. $= 4.04 \times 10^{-12}$

Table 5–2 (*cont.*): Frequency distributions for *the* and *too* in scenes of at least 1000 words

	Shakespeare		Fletcher			
	$\overline{x}$	$s_{\overline{x}}$	$\overline{x}$	$s_{\overline{x}}$	t'	df
a	17.77	0.475	20.50	0.675	3.31	109.2
all	4.39	0.153	7.41	0.429	6.62	66.8
and	29.16	0.507	32.07	1.139	2.33	74.8
are	4.06	0.167	6.46	0.344	6.27	79.1
as	7.10	0.234	5.92	0.426	-2.41	86.7
by	4.56	0.169	3.23	0.269	-4.20	97.3
dare	0.24	0.038	1.23	0.138	6.92	61.1
did	2.10	0.137	0.86	0.125	-6.65	174.7
do	4.78	0.175	6.03	0.352	3.20	80.6
in	13.39	0.313	8.55	0.407	-9.43	120.0
must	1.75	0.117	3.41	0.294	5.24	70.2
no	4.73	0.156	6.18	0.325	4.05	78.5
now	3.09	0.131	5.34	0.368	5.78	66.8
of	19.83	0.518	12.74	0.533	-9.53	155.1
so	6.44	0.196	5.33	0.348	-2.77	88.7
sure	0.36	0.040	1.58	0.189	6.32	57.8
that	13.57	0.273	12.39	0.517	-2.01	84.1
the	32.31	0.719	22.82	0.868	-8.42	130.2
these	1.49	0.103	2.82	0.331	3.82	63.4
to	22.26	0.383	19.50	0.533	-4.20	111.6
too	1.72	0.105	4.52	0.338	7.90	63.4
which	2.87	0.155	1.24	0.152	-7.53	162.5
with	8.58	0.230	6.45	0.348	-5.10	102.5

For the 23 marker words selected at this stage, the sample mean and standard error are shown for each author. The statistic t' and the number of degrees of freedom are also given.

Table 5–3: Potential marker words and some statistics

5.2 Some Frequent Word Classes

Mosteller and Wallace relied on individual words in their examination of *The Federalist* papers. They note that variables based on grammatical concepts are attractive but do not pursue this idea very far. Possibly they were discouraged by the problems encountered in earlier research by Mosteller and Williams, who made counts of nouns and adjectives and were "appalled" by the difficulties that arose in classifying words into these classes [113, p. 9]. While one of the general conclusions of Mosteller and Wallace's analyses is that pronouns and verbs are often affected by context, they do not consider counting "any pronoun" or "forms of *have*" as individual variables.

A full examination of words and grammatical class must await the complete tagging of a large number of 16th and 17th century texts. Software for grammatical analysis of English is being developed, and within the next decade a complete study of grammatical classes and structures in the Shakespeare and Fletcher canons may be feasible. However, some initial efforts may prove useful for the purposes of this study. Nouns and adjectives are numerous and impossible to enumerate completely, but the forms of several closed grammatical classes can be exhaustively listed and counted as a group. These include pronouns and some modal verbs. These are examined in this section along with all the forms of other common verbs. Some of these may prove useful discriminators as a group even though none of the individual forms are effective enough to be considered on their own.

5.2.1 Pronouns

Working from Barber's discussion of pronouns in *Early Modern English* [7], every pronoun form used in the early 17th century was counted in the control plays. The counts and rates for each form are listed in Table 5–4. Ignoring the well-known differences involving *you, ye, them* and *'em,* Fletcher's rate for

	Fletcher		Shakespeare	
	Count	Rate	Count	Rate
I	4518	33.93	12033	28.24
me	1554	11.67	4229	9.92
my	1386	10.41	6101	14.32
mine	164	1.23	588	1.38
thou	575	4.32	2694	6.32
thee	329	2.47	1515	3.55
thy	296	2.22	1765	4.14
thine	26	0.19	195	0.46
he	984	7.39	3630	8.52
him	789	5.92	2551	5.99
his	677	5.08	3234	7.59
she	609	4.57	1403	3.29
her	864	6.49	2078	4.88
hers	4	0.03	30	0.07
it	2172	16.31	5600	13.14
its	0	0.00	6	0.01
we	683	5.13	1959	4.60
vs	427	3.21	1037	2.43
our	370	2.78	1586	3.72
ours	12	0.09	47	0.11
you	2036	15.29	7672	18.00
ye	1925	14.46	120	0.28
your	1527	11.47	3514	8.25
yours	37	0.28	156	0.37
they	632	4.75	1216	2.85
them	41	0.31	1032	2.42
'em	482	3.62	37	0.09
their	358	2.69	972	2.28
theirs	5	0.04	21	0.05
TOTAL	23482	176.34	67021	157.28

Table 5–4: Counts and rates for pronouns in the control set

most pronouns is higher than Shakespeare's. Summing the counts and rates reveals that Fletcher uses pronouns at a rate of 174.0 per thousand compared to Shakespeare's 155.5. The different rates in the two playwrights' works and the high overall frequency of pronouns mean that this class has excellent potential as a discriminator. In fact, t tests conducted later in the analysis produce a value of $t' = 6.7$ when pronouns are measured in scenes of at least 1000 words.

Forms from all four cases were grouped: nominative, accusative, possessive and determiner-pronouns (as Barber calls them) such as *my*. It would have been desirable to study each case on its own, but parsing would be required to determine the case of some forms such as *it* and *her*. Since suitable parsing software does not exist the only solution (other than parsing, for example, every *it* in 34 plays by hand) was to treat all cases as one group.

One interesting point about the *them* and *'em* distinction was quickly noted. If one considers the two as surface realizations of the same grammatical form, then Fletcher's combined rate for the two is 1.6 times that of Shakespeare. This difference results in a significant t' statistic, but all pronouns taken as a group appear to be a more powerful variable.

5.2.2 Some Common Verbs

All the paradigmatic forms for a number of common verbs were examined. Table 5–5 lists the count and rate for each form of *to be*. Two forms *are* and *being* were noted at an early stage of the study because of their large distinctiveness ratios. The first person form *am* occurs more frequently in Fletcher; this may correspond to a more frequent use of *I* in his plays than in Shakespeare's (but as noted earlier Fletcher uses most pronouns more often than the elder playwright). The pooled rates for forms of *to be* are almost identical, and little is to be gained from using "forms of *to be*" as a single variable. The two individual forms *are* and *being* were retained for further examination.

	Fletcher		Shakespeare	
	Count	Rate	Count	Rate
am	521	3.913	1222	2.867
are	871	6.541	1790	4.199
art	108	0.811	442	1.037
be	1138	8.546	3595	8.434
been	116	0.871	384	0.901
beest	8	0.060	11	0.026
being	32	0.240	357	0.838
is	2029	15.237	6972	16.356
was	331	2.486	1252	2.937
wast	4	0.030	25	0.059
were	231	1.735	921	2.161
wert	8	0.060	37	0.087
TOTAL	5397	40.53	17008	39.902

Table 5–5: Counts and rates for forms of *to be* in the control set

	Fletcher		Shakespeare	
	Count	Rate	Count	Rate
did	107	0.804	893	2.095
didst	12	0.090	91	0.213
do	801	6.015	1961	4.600
does	56	0.421	160	0.375
doing	12	0.090	44	0.103
done	149	1.119	332	0.779
dost	48	0.360	185	0.434
doth	12	0.090	463	1.086
TOTAL	1197	8.989	4129	9.685

Table 5–6: Counts and rates for forms of *to do* in the control set

	Fletcher		Shakespeare	
	Count	Rate	Count	Rate
had	310	2.33	749	1.76
hadst	11	0.08	41	0.10
has	297	2.23	188	0.44
hast	88	0.66	300	0.70
hath	32	0.24	1036	2.43
haue	1145	8.60	3132	7.36
hauing	13	0.10	73	0.17
TOTAL	1896	14.24	5524	12.96

Table 5–7: Counts and rates for forms of *to have* in the control set

The verb *to do* has been used in several ways to support the division of *Henry VIII*. Fletcher rarely uses the third person form *doth,* preferring the newer form *does.* Partridge evaluates this usage in the play and also notes that the variation in the number of "expletive" forms of auxiliary *do* (where no emphasis is intended) corresponds to Spedding's division [120]. Table 5–6 lists all the paradigmatic forms of the verb *to do*. The difference in the rates of the preterite form *did* was recognized earlier due to the large distinctiveness ratio. The form *do* also shows a significant difference, but the t' value of 3.2 (with 81 degrees of freedom, measured in scenes of at least 1000 words) is not as high as some other markers found.

As in the case of *them* and *'em,* combining the counts for *does* and *doth* reveals an interesting difference. Fletcher uses either form only a third as often as Shakespeare, although the combined rates are fairly low. This may be tied to differences in the use of auxiliary *do*. (Barber describes the use of *do* as an auxiliary in Early Modern English [7, pp. 263–267]). A full linguistic analysis of *to do* might uncover some useful information regarding each author's style. But because the rates for individual forms of *to do* are low, this would probably not yield information that could be analyzed statistically to help decide the authorship of small scenes.

Shakespeare and Fletcher exhibit the same preferences regarding third person endings for another auxiliary verb: *to have.* Again, this characteristic has been used in previous linguistic analysis of their texts. From examination of the lists of words with large distinctiveness ratios, *hath* is 10.2 times more common in Shakespeare's texts, while *has* is 5.0 times more frequent in the Fletcher control texts. Table 5–7, which lists all the paradigmatic forms, shows that the combined rate for the two forms is about the same in either author, unlike the pair *does/doth.* Like *to do,* the combined rate for all forms of *to have* is similar in both writers and thus cannot be used for discrimination.

5.2.3 Modal Verbs

Barber classifies auxiliary verbs in Early Modern English into two classes, primaries and modals [7, pp. 253–260]. Modals are used with lexical verbs to form a verb phrase, the lexical verb occurring in its base form without a linking *to.* Barber lists 12 main forms of modal verbs:

can	dare	may	mote	shall	will
couthe	durst	might	must	should	would

(By Shakespeare's day *couthe* and *mote* were not commonly used. An alternate form *mought* is sometimes used for *might.* None of these three forms occurs in the 34 plays studied.)

Modal auxiliaries do not have infinitive or *-ing* forms. They also lack the third person inflections *-es* and *-eth.* Such forms for two of these verbs, *dare* and *will,* do occur (for example, *he dareth* and *to will*), but Barber regards these as grammatically distinct lexical verbs. Unfortunately, this means that parsing is required to count modal forms of *will* and *dare* automatically. Since the amount of text involved is considerable, hand parsing was not attempted. (Again, a complete analysis of modals in Jacobean drama must await further development of software for grammatical analysis.)

Forms of two modal verbs have already been recognized as potential markers when common words with large distinctiveness ratios were found. There is only one form of *must,* so in effect the statistics for this modal are included in Table 5-3 (page 193). Like *must,* the base form *dare* is also a Fletcher "plus" word on its own. Another form of this modal, *durst,* is not common but is also favored by Fletcher. His rate of use is 0.28 per thousand compared to Shakespeare's 0.063, a ratio of 4.40. Whether or not the difference in usage for these two forms of *dare* is confined to occurrences of the modal or lexical verb was not determined. But the form *dare* was studied as an individual variable (and eventually plays an important role in the final analyses).[1]

Three other modals remain whose forms can be recognized from their spelling alone. The various paradigmatical forms for *can, may* and *shall* were counted and grouped by base form:

Modal Base Form	Fletcher rate	Shakespeare rate
can	4.09	3.41
may	3.21	2.52
shall	6.54	6.86

The overall rates in the two authors suggest that the rates of these modals are not sufficiently different to be useful in attributing disputed samples.

The study of these common word classes has discovered some potentially useful variables. When counts for personal pronouns are grouped according to grammatical class, the difference in rates between the two dramatists appears to be enough to warrant further study. The overall rates of occurrence for

[1]All paradigmatical forms of *dare* (both lexical and modal) could have been combined and treated as a single variable. This was done using counts from each design-set play, and the between-author difference compared to that for the base form alone. The t test results were very close, with the difference in rate of use for the base form alone slightly more significant. Although combining all the forms would have resulted in a more frequent variable (an average rate of 0.47 compared to 0.28 in the Shakespeare plays; 1.82 compared to 1.59 in Fletcher's), the single form *dare* was used in the remainder of the study, due to the t test result and the effort required for pooling counts for all the forms.

the verbs *be, have* and *do* are almost identical in Shakespeare and Fletcher, although some differences between individual forms or pairs of forms (such as *does* and *doth*) have been noted. Several forms of modal verbs have already been recognized as possible markers. The remaining modals that have forms that can be recognized without parsing do not exhibit a marked difference in the rate of occurrence between these two authors. Word classes may certainly prove important in future authorship studies, but this will depend on the development of successful automatic tagging and parsing software for Early Modern English dramatic texts.

5.3 *Where-/There-* Compounds

During the initial examination of words with high distinctiveness ratios, *therefore* was checked for spelling variants. The form *therfore* was found, and when examining the complete word lists for other variants of words beginning *there-* (such as *thereby* and *thereof*), I noticed that Shakespeare uses a large number of such words in comparison to Fletcher. This is also true for forms beginning with *where-* such as *wherefore* and *whereof.* Table 5–8 (on page 204) lists the forms and rates for both prefixes along with *there* and *where.*

Perhaps Fletcher uses a wider number of these forms in plays that were not examined (only 6 Fletcher plays were used in the control set compared to 20 Shakespeare texts).[2] One of the two Fletcher plays in the test set, *Valentinian,* contains occurrences of *where, there, wherefore* (1 occurrence) and *therefore* (10 occurrences). On the other hand, the other Fletcher test-set play, *Monsieur*

---

[2]Examination of Spevack's one-volume *Harvard Concordance* [155] reveals a number of occurrences in Shakespeare plays that were not examined in this study (including some forms not found in these 26 plays): *thereabout* in *Hamlet; thereafter* in *2 Henry IV; whereas* in *1 Henry VI* (3 occurrences), *2 Henry VI* and *Pericles* (2 occurrences); *whereout* in *Troilus and Cressida;* and *wheresoever* in *Measure for Measure* and *Othello.*

Thomas contains a number of more unusual forms: one occurrence each of *thereabout, thereafter, thereby* and *wherein.* The occurrence of *thereby* is in two lines of a ballad sung at the end of III.iii. Playwrights often made use of popular songs and ballads; this appears to be the case here so this occurrence is probably not noteworthy.

In any case, if the counts for all the forms (excluding the two simple forms *there* and *where*) are combined, the distinctiveness ratio based on the counts in the control set texts is 11.8 in Shakespeare's favor. This ratio is slightly higher than that for *hath,* a Shakespeare "plus" word that has been noted by scholars and used as evidence in studies by Partridge and others. But *hath* is more frequent in Shakespeare (2.43 per thousand) than this *where/there-* group (1.54 per thousand).

The *where/there-* group was not used as a variable in the function word analysis described in Chapter 6, partly because of its relatively low rate of occurrence (especially in Fletcher). In addition, the various forms do not comprise a grammatical class. Orthographically they are very similar, but whether or not this relationship justifies treating them as a group could be debated. However, scenes of *Henry VIII* and *The Two Noble Kinsmen* containing occurrences of these forms are listed in Table 5–9 on page 205. (The attributions of scenes in this table will be discussed further in Chapter 7.) For the most part, these forms occur in scenes agreed to be by Shakespeare. The second part of III.ii of *Henry VIII* contains a single occurrence of *therefore,* but this is not outside the range of Fletcher's use. III.iii of *The Two Noble Kinsmen,* usually assigned to Fletcher, contains an occurrence of *thereby.* While this word is not used by Fletcher in the 8 plays of known authorship, the usage here "and thereby hangs a tale" might be considered a stock phrase. It is also difficult to assess the significance of the occurrence of *wherfore* in III.v ("and do you still cry where, and how, and wherfore") although this word occurs 6.5 times as often in the Shakespeare control plays as in Fletcher's.

Perhaps one might wish to assign more significance to occurrences of the

more unusual forms in scenes which are generally accepted as Fletcher's. I.iii of *Henry VIII* contains two of these forms, *thereunto* and *wherewithall.* Neither word occurs in the 8 Fletcher plays used in this study; *thereunto* does not occur in the 24 Shakespeare texts and *wherewithall* only once (in *Richard II,* V.i.55). These two occurrences in a scene of only 587 words produce a rate of 3.41 per thousand, much higher than that of any Fletcher scene in the control set. The single occurrence of *thereto* in IV.iii of *The Two Noble Kinsmen* again is interesting; the scene is often attributed to Fletcher, but the occurrence of this word is unparalleled in his texts examined in this study. It appears that these scenes may contain Shakespearean characteristics in addition to the Fletcher traits that have led scholars to assign them to the younger dramatist. (The occurrences in these two scenes will be discussed further in Chapter 7 when the two plays are examined.)

As noted above, these forms are infrequent (even in Shakespeare), and they were not used in a more statistically rigorous analysis. A single occurrence in a scene might simply echo another play or originate in a popular song or the source for a play. But some of the occurrences in the disputed plays could be used as evidence that Shakespeare may be responsible for the vocabulary of some scenes attributed to Fletcher. The discovery of this difference in usage is an excellent illustration of the power of computers in the recognition of lexical characteristics that may not be discovered through traditional study.

Fletcher			Shakespeare		
	Count	Rate		Count	Rate
there	511	3.838	there	1192	2.796
therefore	9	0.068	therefore	345	0.809
thereafter	1	0.008	thereabouts	3	0.007
			thereat	1	0.002
			thereby	13	0.030
			therein	32	0.075
			thereof	14	0.033
			thereon	1	0.002
			thereto	12	0.028
			thereupon	5	0.012
			therevpon	1	0.002
			therewith	1	0.002
			therewithall	7	0.016
where	254	1.907	where	693	1.626
whereabouts	1	0.008	where-about	1	0.002
whereby	2	0.015	whereby	5	0.012
wherefore	3	0.023	wherefore	64	0.150
wherein	1	0.008	wherein	64	0.150
TOTAL	782	5.875	where-euer	1	0.002
			where-vntil	2	0.005
			whereat	3	0.007
			whereof	40	0.094
			whereon	11	0.026
			wheresoere	3	0.007
			wheresomere	1	0.002
			whereto	13	0.030
			whereunto	2	0.005
			whereupon	6	0.014
			wherewith	5	0.012
			wherewithall	1	0.002
			TOTAL	2542	5.959
Excluding *there* and *where*:					
	17	0.13		656	1.535

Table 5–8: Counts and rates for *where/there-* forms in the control set

The Two Noble Kinsmen

Act/Scene	Number of words	Attrib. by*	*where/there-* forms Ct.	Rate
I.i	1821	Sh	4	2.20
I.ii	954	Sh	1	1.05
I.iii	804	Sh	1	1.24
III.i	1051	Sh	2	1.90
III.iii	502	Fl	1	1.99
III.v	1241	Fl	1	0.81
III.vi	2717	Fl	1	0.37
IV.iii	877	Fl?	1	1.14
V.i	1392	Sh†	1	0.72
V.iii	1211	Sh	2	1.65
V.iv	1158	Sh	1	0.86

*Attribution of each scene according to Proudfoot [38] and Hoy [55].
†Except for the first 33 lines (276 words).

Henry VIII

Act/Scene	Number of words	Attributed by* Sped.	Hoy	*where/there-* forms Ct.	Rate
Pro.	268	Fl	?	1	3.73
I.i	1868	Sh	Sh	3	1.61
I.ii	1742	Sh	Sh	4	2.30
I.iii	587	Fl	Fl	2	3.41
II.iv	1924	Sh	Sh	8	4.16
III.iia	1663	Sh	Sh	4	2.40
III.iib	2185	Fl	both	1	0.46
V.i	1507	Sh	Sh	2	1.33

*Attribution of each scene according to Spedding [152] and Hoy [55].

Table 5–9: Occurrences of *where/there-* words in *TNK* and *H8*

5.4 Variation with Date and Genre

An ideal variable in an authorship study will show little or no variance due to an author's conscious stylistic decisions. For a Jacobean dramatist the genre of a play represents such a decision. An ideal variable should also vary as little as possible according to a text's date of composition, since in certain problems the date of a disputed work is unknown. Brainerd has examined these two factors for some words and word classes in Shakespeare's texts. In "Pronouns and Genre in Shakespeare's Drama" [18] he uses discriminant analysis and the analysis of variance and covariance to study rates of pronoun groups. Another study, "The Chronology of Shakespeare's Plays: A Statistical Study" [16], describes how he selected words that vary significantly according to date and used linear regression to re-evaluate the dating of Shakespeare's plays as determined by traditional scholarship.

Several of his results are of interest here. Pronouns showed significant variation in rate according to genre (both the class as a whole and most groups determined by number and person). The rates for several modal verbs showed a significant increase with time. To analyze the relationships between these rates and the two factors genre and time, Brainerd made use of several sophisticated statistical procedures (such as regression on principal components and analysis of covariance). For my study, it is enough to discover which words vary according to these factors and compare the variation to the differences between Shakespeare and Fletcher.

In Section 4.3.1 (page 163) the analysis of variance procedure was used to determine if the rate of occurrence of collocations and proportional pairs varies when Shakespeare's plays are grouped by date of composition or by genre. The section presents results of similar tests for the 23 individual marker words and personal pronouns. To test for changes during Shakespeare's career, the 20 plays were divided into five groups corresponding to date of composition. Likewise the

plays were grouped into comedies, tragedies, histories and romances to determine if word rates differ in plays of different genre.[3]

Word rates were measured in acts, and a mean rate and standard error were determined for each group. Fisher's F and an associated probability were calculated; probability values less than 5% indicate a significant variation between the different groups. These statistics for date of composition are listed in Table 5–10. The results for grouping by genre are listed in Table 5–11. (These tables also include each word's mean rate and standard error calculated from all the acts of each author, plus a row of t statistics for the groups that will be explained later.)

A glance at the probabilities in these tables reveals that a large number of the marker words vary significantly between groups. For the five period-groups, 15 of the 25 variables have F values significant at the 5% level; when grouped by genre, 11 of the variables show significant variation. Again, as noted in Section 4.3.1, the ANOVA procedure assumes that the variances for the groups are equal. Tests to validate this are sensitive to departures from normality. Since there are good reasons for expecting word-rate variables to be non-normal, such tests were not carried out. While it would be unwise to accept each individual probability without reservation, in many cases the values of F are extremely large, and it is clear from inspection that the group rates are not the same.

[3]While the names of the plays in each group are listed in full in Section 4.3.1, abbreviations of the members are listed here. The plays in the date of composition groups are: (1) *CE*, *LLL*, *TGV* and *TS*; (2) *MND*, *Rom*, *KJ* and *R2*; (3) *MV*, *1H4*, *MAN*, *H5* and *JC*; (4) *MWW*, *TN* and *AWW*; and (5) *Mac*, *Cor*, *Cym* and *WT*. The comedies are *CE*, *LLL*, *TGV*, *TS*, *MND*, *MV*, *MAN*, *MWW*, *TN* and *AWW*. The tragedies are *Rom*, *JC*, *Mac* and *Cor*. The histories are *KJ*, *R2*, *1H4* and *H5*. The romances are *Cym* and *WT*.

Word		Group Statistics					Overall	
F	prob.	$\overline{x}_1$	$\overline{x}_2$	$\overline{x}_3$	$\overline{x}_4$	$\overline{x}_5$	$\overline{x}_{Sh}$	$\overline{x}_{Fl}$
		$s_{\overline{x}_1}$	$s_{\overline{x}_2}$	$s_{\overline{x}_3}$	$s_{\overline{x}_4}$	$s_{\overline{x}_5}$	$s_{\overline{x}_{Sh}}$	$s_{\overline{x}_{Fl}}$
		t_1	t_2	t_3	t_4	t_5		
a		19.45	15.35	18.27	19.57	14.51	17.39	20.43
6.52	0.11×10^{-3}	1.19	0.77	0.92	0.84	0.51	0.45	0.63
		[0.73]	5.11	[1.94]	[0.82]	7.30		
all		4.29	4.65	4.67	3.47	4.54	4.38	7.73
2.46	0.51×10^{-1}	0.35	0.31	0.28	0.28	0.20	0.13	0.32
		7.25	6.91	7.20	10.02	8.45		
and		28.72	31.68	31.95	26.37	26.21	29.26	32.25
7.84	0.16×10^{-4}	1.13	0.87	1.05	0.92	0.77	0.50	1.20
		2.14	[0.38]	[0.19]	3.89	4.24		
are		4.06	3.38	4.45	3.97	4.86	4.17	6.48
3.40	0.12×10^{-1}	0.35	0.36	0.27	0.25	0.22	0.14	0.33
		5.03	6.35	4.76	6.06	4.08		
as		6.79	6.52	7.48	6.59	7.79	7.07	5.73
1.46	0.22	0.53	0.34	0.44	0.43	0.54	0.21	0.38
		[-1.63]	[-1.55]	-3.01	[-1.50]	-3.12		
by		4.91	4.30	4.71	4.30	4.07	4.48	3.42
1.29	0.28	0.41	0.26	0.24	0.30	0.29	0.14	0.29
		-2.97	-2.26	-3.43	-2.11	[-1.58]		
dare		0.20	0.14	0.34	0.15	0.42	0.26	1.27
2.87	0.27×10^{-1}	0.08	0.05	0.07	0.07	0.08	0.03	0.14
		6.64	7.60	5.94	7.16	5.27		
did		2.08	1.99	2.76	1.69	2.15	2.18	0.80
1.49	0.21	0.37	0.20	0.40	0.32	0.24	0.15	0.09
		-3.36	-5.43	-4.78	-2.68	-5.27		
do		3.69	4.69	5.07	5.16	4.47	4.60	5.92
2.70	0.35×10^{-1}	0.32	0.47	0.35	0.34	0.25	0.17	0.28
		5.24	2.25	[1.90]	[1.73]	3.86		
in		13.07	13.64	14.62	14.03	12.00	13.50	8.60
2.93	0.25×10^{-1}	0.71	0.56	0.56	0.63	0.49	0.28	0.30
		-5.80	-7.93	-9.48	-7.78	-5.92		
must		1.52	2.15	1.71	1.62	2.14	1.83	3.29
1.95	0.11	0.25	0.20	0.17	0.19	0.22	0.10	0.22
		5.32	3.83	5.68	5.74	3.70		
no		4.59	4.35	4.75	4.80	4.81	4.66	6.25
0.37	0.83	0.38	0.31	0.26	0.40	0.25	0.14	0.31
		3.38	4.33	3.71	2.87	3.62		

The first column of statistics corresponds to the earliest plays; the fifth column to the latest. Statistics are for rates per 1000 words.

Table 5–10: Word-rate ANOVA results in Shakespeare by period of composition

Word F	prob.	Group Statistics $\overline{x}_1$ $s_{\overline{x}_1}$ t_1	$\overline{x}_2$ $s_{\overline{x}_2}$ t_2	$\overline{x}_3$ $s_{\overline{x}_3}$ t_3	$\overline{x}_4$ $s_{\overline{x}_4}$ t_4	$\overline{x}_5$ $s_{\overline{x}_5}$ t_5	Overall $\overline{x}_{Sh}$ $s_{\overline{x}_{Sh}}$	 $\overline{x}_{Fl}$ $s_{\overline{x}_{Fl}}$
now		3.55	3.48	3.00	3.37	3.01	3.27	5.92
1.06	0.38	0.28	0.32	0.18	0.30	0.24	0.12	0.42
		4.70	4.62	6.39	4.94	6.02		
of		15.25	20.55	22.18	20.01	19.83	19.63	12.73
7.01	0.54×10^{-4}	1.02	1.36	1.01	0.82	0.52	0.51	0.54
		-2.18	-5.34	-8.25	-7.41	-9.47		
Pronouns		166.02	141.15	152.03	174.45	158.19	157.33	176.49
10.25	0.59×10^{-6}	5.20	2.80	3.39	3.72	3.17	1.99	2.32
		1.84	9.72	5.95	0.47	4.66		
so		6.20	6.16	6.37	5.81	6.71	6.28	5.14
0.57	0.69	0.52	0.38	0.26	0.41	0.44	0.18	0.25
		-1.84	-2.24	-3.41	-1.40	-3.10		
sure		0.47	0.13	0.43	0.49	0.34	0.37	1.63
2.81	0.30×10^{-1}	0.14	0.05	0.05	0.12	0.05	0.04	0.15
		5.65	9.49	7.59	5.93	8.16		
that		14.78	13.42	13.35	13.44	13.48	13.70	11.80
1.00	0.41	0.68	0.73	0.53	0.75	0.41	0.28	0.55
		-3.41	-1.77	-2.03	-1.76	-2.45		
the		28.08	32.04	34.21	28.65	37.35	32.30	23.45
6.21	0.17×10^{-3}	1.71	1.46	1.64	1.16	1.28	0.75	0.77
		-2.47	-5.20	-5.94	-3.73	-9.31		
these		1.61	2.02	1.41	1.06	1.49	1.54	3.01
2.52	0.46×10^{-1}	0.27	0.27	0.15	0.13	0.15	0.09	0.27
		3.67	2.59	5.18	6.51	4.92		
to		23.47	22.98	20.44	20.56	24.28	22.35	19.40
5.69	0.37×10^{-3}	0.97	0.86	0.53	0.62	0.56	0.36	0.46
		-3.79	-3.67	-1.48	-1.50	-6.73		
too		1.76	1.45	1.58	1.59	1.75	1.63	4.40
0.49	0.74	0.21	0.22	0.15	0.16	0.17	0.08	0.28
		7.54	8.28	8.88	8.71	8.09		
which		2.20	2.78	2.59	2.11	4.53	2.86	1.23
9.87	0.99×10^{-6}	0.37	0.26	0.26	0.38	0.28	0.16	0.12
		-2.49	-5.41	-4.75	-2.21	-10.83		
with		8.39	9.87	8.69	7.68	8.52	8.68	6.72
3.27	0.15×10^{-1}	0.48	0.45	0.31	0.47	0.40	0.19	0.21
		-3.19	-6.34	-5.26	-1.86	-3.98		

The first column of statistics corresponds to the earliest plays; the fifth column to the latest. Statistics are for rates per 1000 words.

Table 5–10 (*cont.*): Word-rate ANOVA results in Shakespeare by period of composition

Word			Group Statistics				Overall	
	F	prob.	$\overline{x}_1$	$\overline{x}_2$	$\overline{x}_3$	$\overline{x}_4$	$\overline{x}_{\mathrm{Sh}}$	$\overline{x}_{\mathrm{Fl}}$
			$s_{\overline{x}_1}$	$s_{\overline{x}_2}$	$s_{\overline{x}_3}$	$s_{\overline{x}_4}$	$s_{\overline{x}_{\mathrm{Sh}}}$	$s_{\overline{x}_{\mathrm{Fl}}}$
			t_1	t_2	t_3	t_4		
a			19.49	14.13	16.15	15.65	17.39	20.43
	10.69	0.39×10^{-5}	0.61	0.69	0.92	0.74	0.45	0.63
			1.07	6.74	3.84	4.92		
all			4.07	4.59	5.01	4.32	4.38	7.73
	2.71	0.49×10^{-1}	0.19	0.28	0.32	0.26	0.13	0.32
			9.83	7.38	6.01	8.27		
and			28.32	28.73	34.52	24.63	29.26	32.25
	15.70	0.21×10^{-7}	0.63	0.89	0.95	0.51	0.50	1.20
			2.90	2.36	-1.48	5.84		
are			4.20	4.71	3.34	4.57	4.17	6.48
	3.82	0.12×10^{-1}	0.20	0.33	0.28	0.30	0.14	0.33
			5.91	3.79	7.26	4.28		
as			6.78	6.54	7.46	8.85	7.07	5.73
	3.58	0.17×10^{-1}	0.29	0.47	0.41	0.70	0.21	0.38
			-2.20	-1.34	-3.09	-3.92		
by			4.63	3.94	4.46	4.87	4.48	3.42
	1.50	0.22	0.22	0.27	0.24	0.35	0.14	0.29
			-3.32	-1.31	-2.76	-3.19		
dare			0.21	0.34	0.23	0.38	0.26	1.27
	1.22	0.31	0.05	0.07	0.05	0.13	0.03	0.14
			7.13	5.94	7.00	4.66		
did			2.03	2.63	2.26	1.92	2.18	0.80
	0.88	0.45	0.22	0.32	0.33	0.27	0.15	0.09
			-5.17	-5.51	-4.27	-3.94		
do			4.67	5.15	4.15	4.04	4.60	5.92
	1.64	0.19	0.26	0.35	0.32	0.34	0.17	0.28
			3.27	1.72	4.16	4.27		
in			13.82	12.74	14.71	10.91	13.50	8.60
	5.49	0.16×10^{-2}	0.38	0.39	0.74	0.57	0.28	0.30
			-10.78	-8.41	-7.65	-3.59		
must			1.67	2.04	1.93	2.06	1.83	3.29
	1.04	0.38	0.14	0.22	0.18	0.33	0.10	0.22
			6.21	4.02	4.78	3.10		
no			4.86	4.53	4.04	5.12	4.66	6.25
	2.11	0.10	0.22	0.25	0.29	0.25	0.14	0.31
			3.66	4.32	5.21	2.84		

Groups 1–4 correspond to plays as follows: (1) Comedies, (2) Tragedies, (3) Histories and (4) Romances. Statistics are for rates per 1000 words.

Table 5–11: Word-rate ANOVA results in Shakespeare by genre

Word F	prob.	Group Statistics $\overline{x}_1$ $s_{\overline{x}_1}$ t_1	 $\overline{x}_2$ $s_{\overline{x}_2}$ t_2	 $\overline{x}_3$ $s_{\overline{x}_3}$ t_3	 $\overline{x}_4$ $s_{\overline{x}_4}$ t_4	Overall $\overline{x}_{Sh}$ $s_{\overline{x}_{Sh}}$	 $\overline{x}_{Fl}$ $s_{\overline{x}_{Fl}}$
now		3.32	3.09	3.33	3.21	3.27	5.92
0.21	0.89	0.18	0.24	0.21	0.40	0.12	0.42
		5.69	5.85	5.52	4.67		
of		17.62	18.21	26.01	19.94	19.63	12.73
22.55	0.37×10^{-10}	0.61	0.75	0.90	0.88	0.51	0.54
		-6.00	-5.93	-12.65	-6.98		
Pronouns		165.43	151.90	139.81	161.98	157.33	176.49
11.25	0.21×10^{-5}	2.87	2.80	3.28	4.21	1.99	2.32
		3.00	6.76	9.13	3.02		
so		6.07	6.04	6.21	7.98	6.28	5.14
3.60	0.16×10^{-1}	0.28	0.32	0.28	0.48	0.18	0.25
		-2.48	-2.22	-2.85	-5.25		
sure		0.45	0.32	0.17	0.40	0.37	1.63
2.59	0.57×10^{-1}	0.07	0.06	0.04	0.07	0.04	0.15
		7.13	8.11	9.40	7.43		
that		13.82	13.95	13.19	13.61	13.70	11.80
0.31	0.82	0.44	0.45	0.66	0.63	0.28	0.55
		-2.87	-3.03	-1.62	-2.16		
the		29.97	34.24	35.39	34.12	32.30	23.45
3.61	0.16×10^{-1}	1.05	1.77	1.57	1.55	0.75	0.77
		-5.01	-5.59	-6.83	-6.17		
these		1.45	1.59	1.65	1.65	1.54	3.01
0.30	0.83	0.16	0.17	0.17	0.25	0.09	0.27
		4.97	4.45	4.26	3.70		
to		21.74	22.91	22.40	24.24	22.35	19.40
1.61	0.19	0.55	0.69	0.82	0.59	0.36	0.46
		-3.26	-4.23	-3.19	-6.47		
too		1.64	1.66	1.38	2.03	1.63	4.40
1.48	0.23	0.11	0.20	0.19	0.25	0.08	0.28
		9.17	7.96	8.92	6.31		
which		2.32	3.39	2.77	4.74	2.86	1.23
8.97	0.27×10^{-4}	0.22	0.33	0.27	0.35	0.16	0.12
		-4.35	-6.15	-5.21	-9.49		
with		8.48	8.93	9.35	7.86	8.68	6.72
1.70	0.17	0.28	0.47	0.43	0.30	0.19	0.21
		-5.03	-4.29	-5.50	-3.11		

Groups 1–4 correspond to plays as follows: (1) Comedies, (2) Tragedies, (3) Histories and (4) Romances. Statistics are for rates per 1000 words.

Table 5–11 (*cont.*): Word-rate ANOVA results in Shakespeare by genre

As noted in Section 4.3.1, date of composition and genre are often closely related, making it difficult to say which factor might be responsible for the change. However, some of the variations are quite interesting stylistically. Shakespeare's comedies seem to be characterized by high rates for pronouns and the indefinite article *a*, while *the* occurs less frequently in this group than in the other genre groups. The tragedies are marked by a low rate of *a*. Personal pronouns occur at very low rates in the history plays (as noted by Brainerd), while *in*, *of* and *and* have high rates of occurrence. *In* occurs infrequently in the romances, while *so* is much more frequent in this group than in the other three.

Some of these results are also evident in certain period-of-composition groups which are composed for the most part of plays of a particular genre. For example, each play in period-groups 1 and 4 is a comedy, and these two groups also have a high occurrence rate for *a*. Some period-groups are marked by a high or low rate that seems to be independent of the plays' genre. Shakespeare's last plays (*Mac*, *Cor*, *Cym* and *WT*) are marked by a high rate for *the*.

Clearly the use of many of these 24 variables is affected by the influence of genre and date of composition. In some cases the variation is quite large. For example, one frequent marker word that was retained despite having a low t' value was *and* ($t' = 2.33$). The probability associated with the F value for *and* is 0.21×10^{-7}. Although the overall rate of occurrence is greater in Fletcher than Shakespeare, the high rate in Shakespeare's history plays is greater than the overall Fletcher rate. Thus *and* appears to be a good marker of genre in Shakespeare, and one might be willing to use it in an authorship study if the genre of the disputed samples was easily recognized. *Henry VIII* is certainly a history play, but it resembles the romances in many ways. (The overall rate for *and* in the play is 28.5; therefore if it were Shakespeare's unaided work, it would not resemble his other histories in this feature.)

The group rates for some words vary significantly within Shakespeare but still differ considerably from the overall Fletcher rate. For example, the ANOVA results for the preposition *in* are significant when the plays are grouped by

period and genre, but the rate of occurrence for any of the groups is considerably higher than the overall Fletcher rate of 8.60. (The lowest rate for a Shakespeare group is the romances' rate of 10.91.) While significant variations have been discovered for some words, it may still be possible to use them to discriminate between Shakespeare and Fletcher if these internal differences are smaller than the differences between the two writers.

To determine objectively which words should be eliminated and which kept, the t test was used to compare the mean rate of occurrence for each group (either by period of composition or genre) with the overall Fletcher rate. For each word in the tables, the t' statistics are listed in a third row under the mean rates and their standard errors. This value is enclosed in a box when it is less than 2.0, which indicates (for a large number of degrees of freedom) that there is a probability of at least 5% that the population mean for that group is equal to Fletcher's true mean rate. (The value for *and* in Shakespeare's histories is also so marked, since it is higher than Fletcher's rate which is in turn higher than the overall Shakespeare rate.) Any variable was eliminated from the set of markers when any of its groups had a t' value less than 2.0.

This procedure eliminated many of the frequent words (for example *a, and, as* and *to*) that were not rejected earlier despite their low t' values. In addition, the pooled set of personal pronouns, which looked as if it might be a valuable marker of authorship, was also eliminated. Pronouns had shown great promise, having a t' value of 6.7 and a rate of occurrence of over 150 in both authors. However, the rate of occurrence in the fourth period-group of plays (composed of the acts of *MWW*, *TN* and *AWW*) almost equals the overall Fletcher rate. Among the words retained, a number do vary significantly within the Shakespeare canon when acts are grouped by genre or period of composition. These words include *are, dare, in, of, sure, the* and *which.*

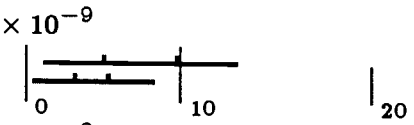
5.5 The Final Set of Marker Words

After eliminating words which have values for Shakespearean period or genre groups that are too close to the overall Fletcher rate, a total of 14 words remain. Table 5–12 contains the same type of graphs used in the last chapter to show the mean rates $\overline{x}$, standard deviations s and interquartile range for these 14 words (counted in scenes of at least 1000 words). Examination of these graphs shows that the bars (representing the interval $\overline{x} \pm 2s$) still overlap considerably for most words. The interquartile ranges overlap very little or not all for several of the words with large t' values (*all, in, of, the* and *too*).

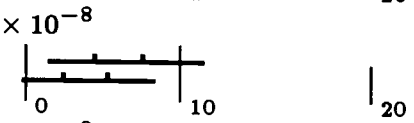
The degree of overlap shown in these graphs was anticipated after examination of the frequency distributions for the same data (some of which were listed earlier, beginning on page 190). Table 5–12 supports the conclusion that the best function-word markers will probably not be useful on their own in discriminating between short scenes of Shakespeare and Fletcher. Nevertheless a statistical method that combines the information from several of these variables can correctly assign scenes of known authorship, as will be demonstrated in the next chapter.

Note: The bars on the right indicate $\overline{x} \pm 2s$. The ticks on top of the bars mark the 1st and 3rd quartiles.

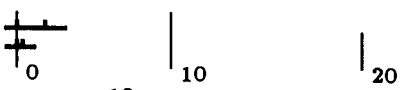
all $t' = 6.62$ $\nu' = 66.82$ $p = 7.63 \times 10^{-9}$
Fl: $\overline{x} =$ 7.41 $s =$ 3.15
Sh: $\overline{x} =$ 4.39 $s =$ 1.98



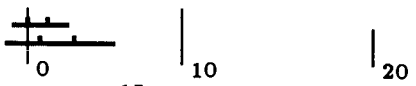
are $t' = 6.27$ $\nu' = 79.12$ $p = 1.79 \times 10^{-8}$
Fl: $\overline{x} =$ 6.46 $s =$ 2.53
Sh: $\overline{x} =$ 4.06 $s =$ 2.16



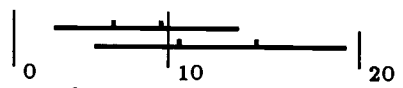
dare $t' = 6.92$ $\nu' = 61.15$ $p = 3.18 \times 10^{-9}$
Fl: $\overline{x} =$ 1.23 $s =$ 1.01
Sh: $\overline{x} =$ 0.24 $s =$ 0.49



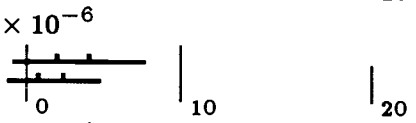
did $t' = -6.65$ $\nu' = 174.68$ $p = 3.98 \times 10^{-10}$
Fl: $\overline{x} =$ 0.86 $s =$ 0.92
Sh: $\overline{x} =$ 2.10 $s =$ 1.78



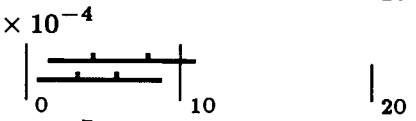
in $t' = -9.43$ $\nu' = 120.04$ $p = 4.88 \times 10^{-15}$
Fl: $\overline{x} =$ 8.55 $s =$ 2.99
Sh: $\overline{x} =$ 13.39 $s =$ 4.06



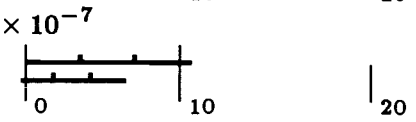
must $t' = 5.24$ $\nu' = 70.22$ $p = 1.62 \times 10^{-6}$
Fl: $\overline{x} =$ 3.41 $s =$ 2.16
Sh: $\overline{x} =$ 1.75 $s =$ 1.51



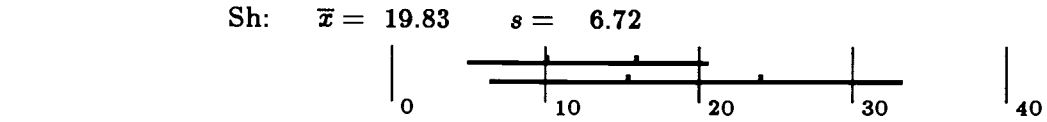
no $t' = 4.05$ $\nu' = 78.45$ $p = 1.20 \times 10^{-4}$
Fl: $\overline{x} =$ 6.18 $s =$ 2.39
Sh: $\overline{x} =$ 4.73 $s =$ 2.02



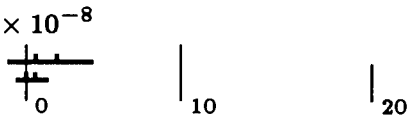
now $t' = 5.78$ $\nu' = 66.80$ $p = 2.22 \times 10^{-7}$
Fl: $\overline{x} =$ 5.34 $s =$ 2.70
Sh: $\overline{x} =$ 3.09 $s =$ 1.70



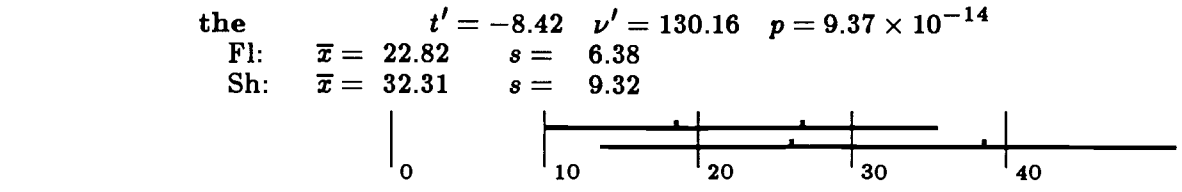
of $t' = -9.53$ $\nu' = 155.08$ $p = 0$
Fl: $\overline{x} =$ 12.74 $s =$ 3.92
Sh: $\overline{x} =$ 19.83 $s =$ 6.72



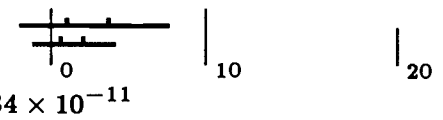
sure $t' = 6.32$ $\nu' = 57.83$ $p = 4.28 \times 10^{-8}$
Fl: $\overline{x} =$ 1.58 $s =$ 1.39
Sh: $\overline{x} =$ 0.36 $s =$ 0.52



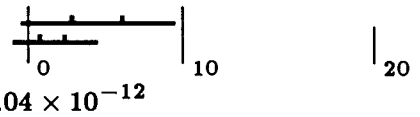
the $t' = -8.42$ $\nu' = 130.16$ $p = 9.37 \times 10^{-14}$
Fl: $\overline{x} =$ 22.82 $s =$ 6.38
Sh: $\overline{x} =$ 32.31 $s =$ 9.32



these $t' = 3.82$ $\nu' = 63.45$ $p = 3.08 \times 10^{-4}$
Fl: $\overline{x} =$ 2.82 $s =$ 2.43
Sh: $\overline{x} =$ 1.49 $s =$ 1.34



too $t' = 7.90$ $\nu' = 63.45$ $p = 5.34 \times 10^{-11}$
Fl: $\overline{x} =$ 4.52 $s =$ 2.49
Sh: $\overline{x} =$ 1.72 $s =$ 1.37



which $t' = -7.53$ $\nu' = 162.52$ $p = 4.04 \times 10^{-12}$
Fl: $\overline{x} =$ 1.24 $s =$ 1.12
Sh: $\overline{x} =$ 2.87 $s =$ 2.00

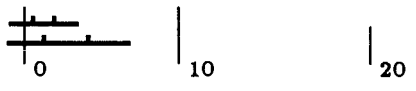


Table 5–12: Internal variation of word-rate variables

5.5.1 Pooled Sets of Infrequent Markers

A number of words with large distinctiveness ratios were eliminated early in the selection process because they were relatively infrequent. For example, *forth* has a distinctiveness ratio of 8.21, but occurs only 7 times in the six Fletcher plays of the control set; it is used by Shakespeare at a rate of 0.43 per thousand. Other words, such as *being* and *still* (see Table 5–1) were considered at the second stage of the selection process but were rejected when t test results suggested they were not among the best discriminators. From a linguistic viewpoint the different rates of use for such words are very interesting; if possible one would like to make some use of them in a statistical analysis.

One way of accomplishing this would be to combine the counts for those infrequent words favored by an author and treat this pooled count as a single variable. This is similar to counting all personal pronouns as a single class. But since the choice of which words are to be treated as a group is not determined by grammar, questions of subjectivity and selection bias arise. Another question regards statistical correlation among members of the pooled set. No statistical method could allow for any relationships between such words once the counts had been combined. If one begins to pool word counts to produce variables with larger frequencies, then one wonders when to stop this process. Taken to its extreme, the rates for all marker words could be combined to produce a single variable which would be used to determine the authorship of a disputed text sample. Out of curiosity this was carried out; the results were not very successful in classifying the test-set scenes. However, this is how Austin (in his study of Greene's *Groats-worth of Wit* [3]) made use of the Chettle and Greene marker words he found in the texts.

Combining measurement values in this way is the idea behind a *linear discriminant function.* Usually the values of the individual measurements (in this case, each single word rate) are weighted to allow for the correlation and relative importance of variables. Such a procedure was demonstrated in the context of

an authorship study by Mosteller and Wallace in their chapter "Weight-Rate Analysis" [113].

With some reservations I decided to combine the rates for a number of words to form what will be called an *infrequent marker pooled set* for each author. This results in another "plus" word-variable for each author. The pooled set of words favored by Shakespeare will be indicated by "Infreq-Sh+" and Fletcher's set by "Infreq-Fl+." The statistical analyses described in the next chapter are carried out twice: once without using these infrequent marker pooled sets and again including these variables. When the rates for "Infreq-Fl+" and "Infreq-Sh+" are used, they are treated just like any of the other 14 words used in the analysis. This in effect limits the contribution of any individual member of either pooled set of words.

Words with large distinctiveness ratios were considered, even if their rate of use was quite low for one of the authors. Only function words were chosen. Complete word lists for the authors were examined to determine if the difference in rate of use might be solely due to spelling variation. When variants were found which did not detract from the distinctiveness ratio, these were included with the primary spelling in the pooled set. For example, the adverb *suddenly* was initially noted in the distinctiveness ratio lists, with a ratio of 29.9 in Fletcher's favor. Variants were found when the complete word lists were examined; these and related forms such as *sudden* and *suddainesse* were included in "Infreq-Fl+." As noted earlier, each of the pairs *them/'em* and *does/doth* is favored by one author or the other when regarded as single grammatical forms. These two pairs of words have therefore been included in these pooled sets.

Table 5–13 (on page 219) lists the words used in the pooled sets along with their overall counts and rates in both authors. At the end of the list of words for each author the total count and rate are listed. In addition, the t' statistics used to compare the two playwrights' rate of use are provided. Comparison of these values with those for the initial set of 23 words considered (Table 5–3 on page 193) shows that these two variables are the best individual discriminators.

The set of words favored by Fletcher is very frequent; the Shakespeare set is moderately so.

Like the other word-rate variables, these two new variables were tested for variation within the Shakespeare texts. The same procedure was used to divide the plays into groups according to date of composition and genre. The results displayed in Table 5–14 show that both markers vary significantly when the ANOVA procedure is applied to the five sets grouped by period of composition; only the pooled set favored by Fletcher varies within Shakespeare when plays are grouped by genre. In fact, this variable "Infreq-Fl+" has a high rate (that is, closer to Fletcher's) in the groups representing Shakespeare's last plays and romances. However, the smallest t statistic for any group is 5.65. Despite the internal variation, the rates for both variables are still different enough from Fletcher's overall rate of use.

Infrequent words favored by Fletcher:

	Fletcher		Shakespeare	
	Count	Rate	Count	Rate
again(e), agen	182	1.367	380	0.892
done	149	1.119	332	0.779
euer	170	1.277	328	0.769
find(e)	170	1.277	286	0.950
nor	208	1.562	455	1.067
off	135	1.014	234	0.549
only, onlie	33	0.248	19	0.045
ready	60	0.451	77	0.181
still	182	1.367	249	0.584
sudden, suddain(e), suddeine	23	0.174	13	0.030
suddenly, suddenlie, suddainly, suddainely	38	0.286	6	0.013
suddainesse	1	0.008	0	0.000
them/'em	523	3.928	1069	2.508
thus	178	1.337	364	0.854
vp	271	2.035	568	1.333
yet	370	2.779	799	1.874
Infreq-Fl+	2693	20.229	5179	12.149

For scenes of at least 1000 words:
$t' = 11.0$, degrees of freedom $= 72.9$, prob. $= 4.00 \times 10^{-15}$

Infrequent words favored by Shakespeare:

	Fletcher		Shakespeare	
	Count	Rate	Count	Rate
being	32	0.240	357	0.838
does	56	0.421	160	0.375
doth	12	0.090	463	1.086
each	9	0.068	100	0.235
forth, foorth	7	0.053	204	0.479
from	273	2.050	1277	2.996
hence	14	0.105	195	0.457
other	50	0.375	304	0.713
rather	29	0.218	197	0.462
self	11	0.083	180	0.422
while	22	0.165	160	0.375
whilst, whilest, whil'st	24	0.181	55	0.129
whom(e)	15	0.113	199	0.467
Infreq-Sh+	554	4.162	3851	9.034

For scenes of at least 1000 words:
$t' = -13.8$, degrees of freedom $= 172.3$, prob. $= 0.0$

Table 5–13: Words in the pooled sets of infrequent markers

ANOVA Results Showing Variation of Infreq-Fl+ and Infreq-Sh+ within Shakespeare by Period of Composition:

Word			Group Statistics					Overall	
	F	prob.	$\overline{x}_1$	$\overline{x}_2$	$\overline{x}_3$	$\overline{x}_4$	$\overline{x}_5$	$\overline{x}_{Sh}$	$\overline{x}_{Fl}$
			$s_{\overline{x}_1}$	$s_{\overline{x}_2}$	$s_{\overline{x}_3}$	$s_{\overline{x}_4}$	$s_{\overline{x}_5}$	$s_{\overline{x}_{Sh}}$	$s_{\overline{x}_{Fl}}$
			t_1	t_2	t_3	t_4	t_5		
Infreq-Fl+			10.75	11.71	11.92	10.61	15.35	12.12	20.24
	9.85	1.02×10^{-6}	0.60	0.52	0.50	0.78	0.66	0.31	0.56
			11.56	11.16	11.08	10.03	5.65		
Infreq-Sh+			9.97	9.70	8.26	7.35	9.76	9.06	4.22
	4.49	2.28×10^{-3}	0.78	0.41	0.43	0.45	0.31	0.24	0.27
			-6.97	-11.16	-7.96	-5.96	-13.48		

ANOVA Results Showing Variation of Infreq-Fl+ and Infreq-Sh+ within Shakespeare by Genre

Word			Group Statistics				Overall	
	F	prob.	$\overline{x}_1$	$\overline{x}_2$	$\overline{x}_3$	$\overline{x}_4$	$\overline{x}_{Sh}$	$\overline{x}_{Fl}$
			$s_{\overline{x}_1}$	$s_{\overline{x}_2}$	$s_{\overline{x}_3}$	$s_{\overline{x}_4}$	$s_{\overline{x}_{Sh}}$	$s_{\overline{x}_{Fl}}$
			t_1	t_2	t_3	t_4		
Infreq-Fl+			10.71	15.00	11.87	14.04	12.12	20.24
	14.56	6.66×10^{-8}	0.35	0.73	0.49	0.85	0.31	0.56
			14.43	5.70	11.25	6.09		
Infreq-Sh+			8.65	9.44	9.34	9.84	9.06	4.22
	1.05	0.374	0.42	0.34	0.43	0.53	0.24	0.27
			-8.87	-12.02	-10.08	-9.45		

Table 5–14: ANOVA results by date and genre for the pooled sets of infrequent markers

5.5.2 Correlation

As noted in Chapter 4 the question of correlation between literary features is important when considering what statistical methods should be used to analyze the data. In that chapter it was seen that, when counted in acts by Fletcher, about 7% of the collocations and proportional pairs tested were significantly correlated. When counted in acts by Shakespeare the proportion was higher, just under 10%. The same procedure can be applied to the 14 individual words and the two pooled sets of infrequent markers.

For all possible pairs of the 16 variables, Kendall's τ, a rank correlation coefficient, was computed using the statistical package SAS. For each pairing the package also determines the probability of a larger degree of correlation. The rates were counted in scenes that were at least 1000 words in length. Table 5–1 lists the combinations that were significant at the 5% level for either author. The total number of possible combinations for the 16 variables is $(16 \times (16 - 1))/2 = 120$. In the Fletcher samples $9/120 = 7.5\%$ of the pairs are significantly correlated. As for collocations and proportional pairs, a larger proportion are significantly correlated in Shakespeare's scenes: $20/120 = 16.7\%$.[4] The larger proportion of significant pairs for Shakespeare does not appear to be due to the fact that there are more Shakespeare samples. When rates from scenes of the 7 latest plays were tested, the number of significant results decreased by 3 to 17.

The only two combinations that are significantly correlated in both author's samples are *in/of* and *are/these*. Perhaps the significance for the second pair is due to relatively frequent occurrences of *these are* or *are these*. The correlation between the two common prepositions is very interesting in its own right and in relation to this study of authorship. *In* and *of* are two of the best individual

---

[4]The Pearson product-moment coefficient ρ was also calculated, but since normality cannot be assumed it is less appropriate here. The number of significant pairings in the Fletcher samples was identical, although the combinations themselves differ slightly. The number of significant pairings in Shakespeare increased by one when ρ was used evaluated at the 5% level.

The following table indicates which word-rate variables are significantly correlated the 5% level of significance with every individual marker word. The results are presented for both writers and are based on rates in scenes of at least 1000 words. The rank correlation coefficient used is Kendall's τ. List items printed in italics indicate a negative correlation.

Marker word	Fletcher	Shakespeare
all	—	are, Infreq-Fl+
are	these	all, sure, these, *Infreq-Sh+*, Infreq-Fl+
dare	no	—
did	*Infreq-Fl+*	*must*, these, *too*, which
in	of, must	of, *Infreq-Fl+*
must	*Infreq-Fl+*, in	*did*
no	dare	*the*
now	—	*the, which*, these
of	*Infreq-Fl+*, in, which	the, in
sure	—	are
the	*too*	of, which, *no, now*
these	are	did, are, now
too	the	*did*
which	of	Infreq-Sh+, the, *now*, Infreq-Fl+, did
Infreq-Fl+	*of, did, must*	*in*, which, all, are
Infreq-Sh+	—	which, *are*

Figure 5–1: Correlated word-rate variables in Fletcher and Shakespeare

markers, but this indicates that results from univariate significance tests of these two words cannot be combined as if independent.

Mosteller and Wallace published a frequency distribution of Pearson correlation coefficients for the rates of 30 words in some of *The Federalist* papers [113, p. 36]. For both Hamilton and Madison, values larger than 0.30 are significant below the 5% level, and inspection shows that 5.4% of the possible word combinations are significant in the Hamilton samples and 3.9% in Madison's. These proportions of correlated pairs are much lower than those for the Shakespeare scenes. One might wonder if this may be due to the fact that the Hamilton and Madison samples are larger. When word rates for the two Jacobeans are

measured in acts rather than scenes of at least 1000 words, the number of significant pairs is roughly the same: 9 for Fletcher and 21 for Shakespeare. In any case, since individual scenes of 1000 words or shorter are disputed, correlation in samples of this length must be taken into account. These results further justify the multivariate discriminant analysis approach outlined in the next chapter.

5.5.3 Minimum Sample Size

Previous research in stylometry has focused some attention on the idea of a minimum sample size: the shortest length of text that can be analyzed by a given method. The authors of "To Couple Is the Custom" express the idea behind this concept: "Language is not random in fine detail and can only be treated as random if the samples are large enough" [102, p. 4]. Merriam also stresses that the determination of a minimum sample size is essential [92, p. 279].

Relatively high or low proportions of occurrences of a literary feature may occur in short samples of text. These might be due to a repetition or stylistic device or simply to random variation. If the rates in a number of such small samples differ a great deal from the mean rate, in either direction, then a large variance may result. When larger samples are taken from the same text (for example, by counting by act instead of by scene) the high and low values often even out and the variance decreases.

Determining a minimum sample size is not a problem in some studies. For example, in confirming or denying the attribution of a novel, the problem is more the degree of variation between a writer's novels unless a novel's integrity is in doubt. In a question of dramatic collaboration, however, one wishes to determine the authorship of sections of a play. Internal variation within a single writer's dramas becomes extremely important. One wants to know how finely a play of known authorship can be sliced up without reaching the stage where some sections begin to look like another author.

Such a lower limit will depend on the nature and frequency of the literary features used in a study, and will probably also depend to some extent on the method of statistical analysis employed. Validation of the statistical procedure on known samples of varying lengths (with the same features used with the disputed text) is the most reliable method of determining the size of the shortest text samples that can be accurately assigned to an author. In past linguistic and stylometric studies, researchers have usually demonstrated that their methods can distinguish one dramatist's plays from another's but have rarely preceded the analysis of a proposed collaboration with a comprehensive analysis of shorter samples from within the undisputed plays.

In the next chapter, before applying statistical methods to the word-rate variables in the two disputed dramas, scenes of known authorship will be evaluated in order to determine the accuracy of the procedure. In this section, the effect of text-sample length on the standard deviation will be examined for some of the 14 individual words. A program was used to count occurrences per block of 250 words in each of the 24 control texts. The means and standard deviations were calculated for these samples, and then counts for contiguous blocks (from the same play) were combined to produce the data for 500 word blocks. This process was repeated until statistics were produced for blocks of 4000 words (about the average size of an act).

Not surprisingly the means for a given word are almost identical when the rates are measured in blocks of different length. Graphs showing the standard deviations of the rate of occurrence of *all, are, dare, in, of, the* and *too* are given in Figure 5–2. Each graph shows that the standard deviation does indeed increase for shorter samples and that this increase becomes more rapid for samples shorter than 1000 words. This increase appears to be less dramatic for infrequent words like *dare* and *too*.

It is difficult to interpret the significance of these graphs in relation to the question: "How short a sample of text can one examine using these word-rate variables?" The observed increase of the standard deviations can be shown to

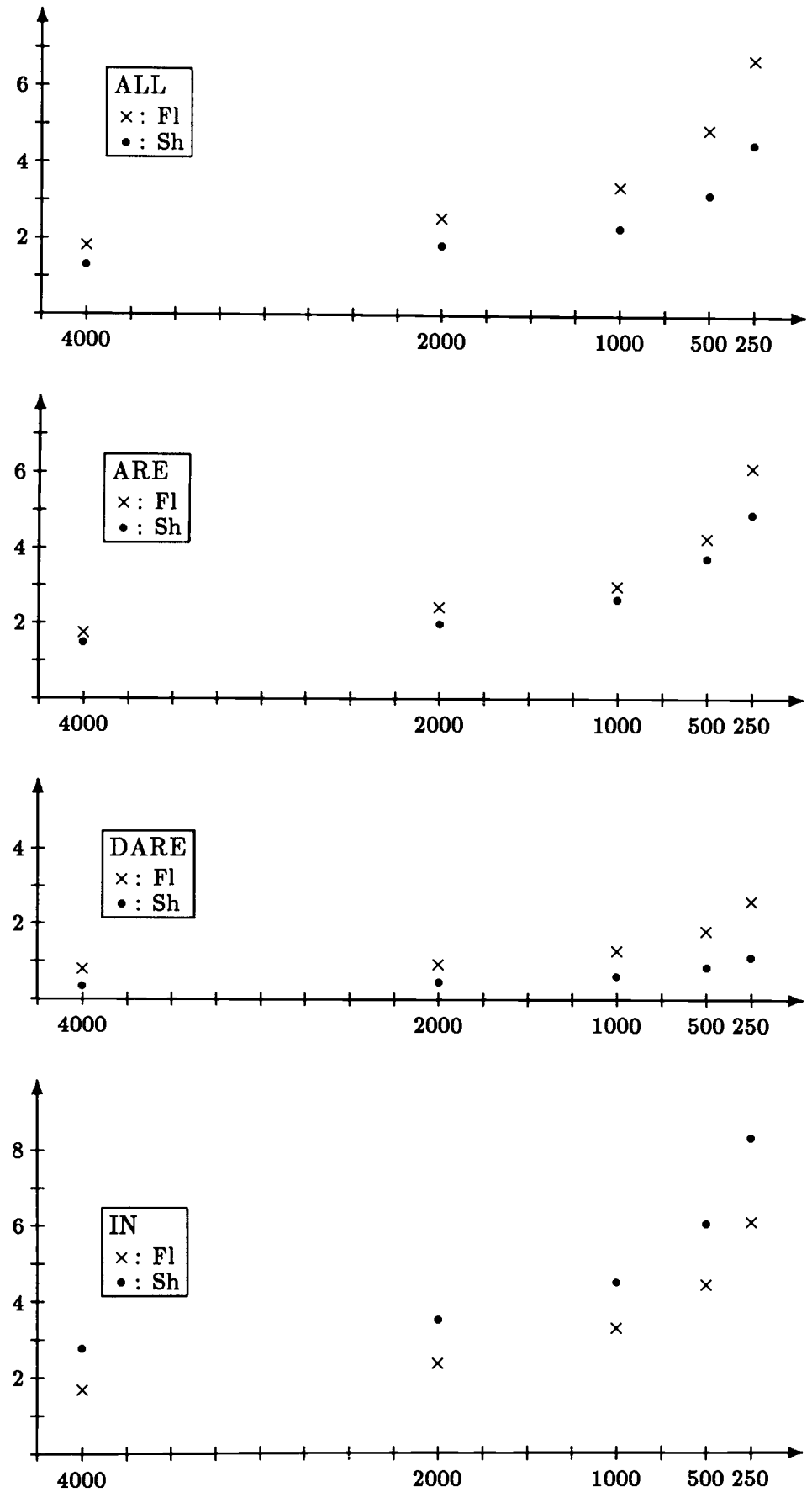


Figure 5–2: Standard deviations in samples of decreasing length

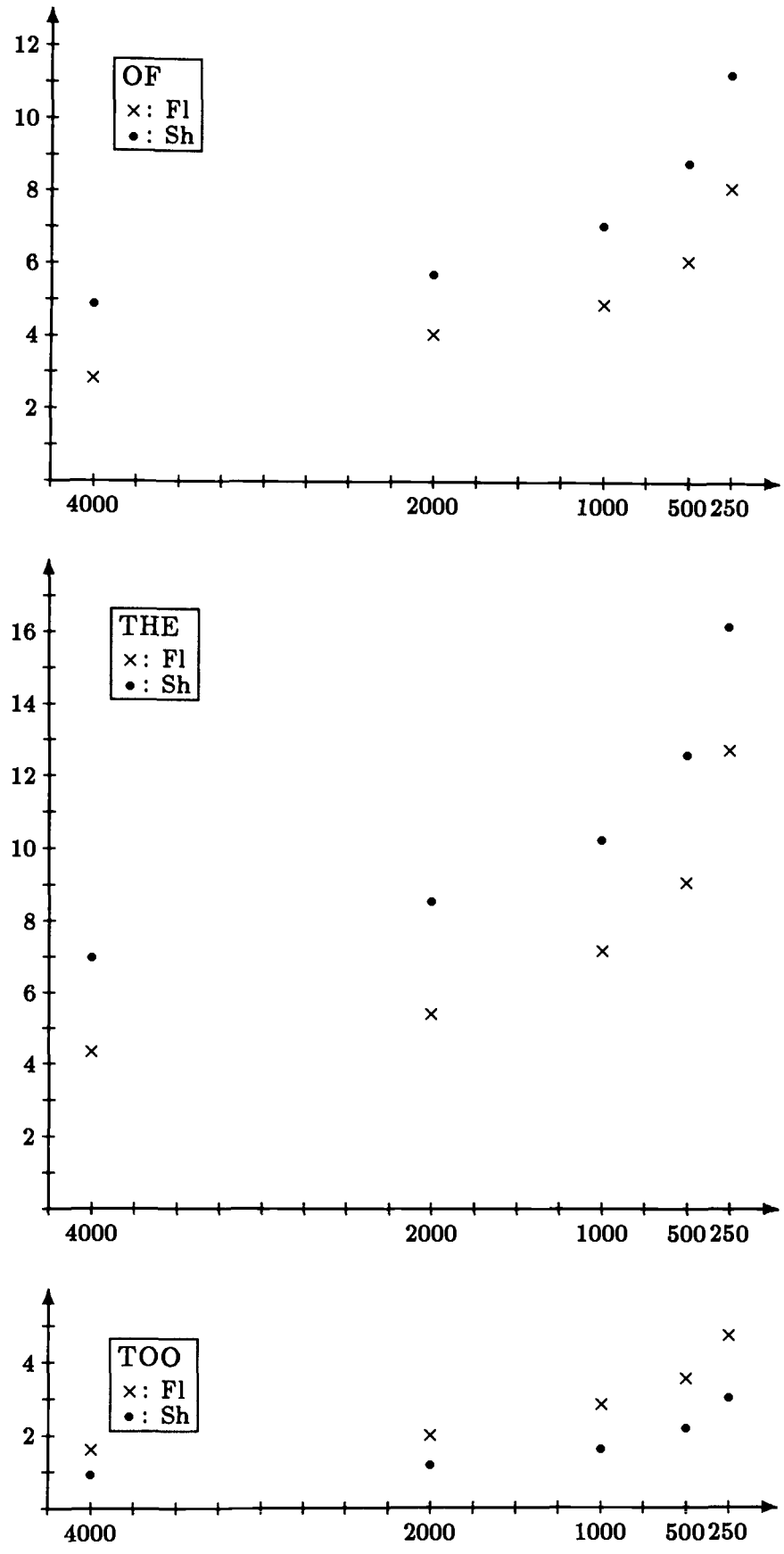


Figure 5–2 (*cont.*): Standard deviations in samples of decreasing length

be a function of the increased variance of the counts underlying the rates, but it is also directly related to the number of samples used in calculating the statistics (and thus the number of words in each block).

Beginning with blocks containing F total words, let x_i represent the number of occurrences in any of the n blocks of this size. The rate of occurrence r_i in a given block is simply x_i/F, with variance:

$$s_r^2 = \frac{1}{n-1}\left[\sum_i (r_i)^2 - \frac{1}{n}\left(\sum_i r_i\right)^2\right]$$

$$= \frac{1}{n-1}\left[\sum_i (x_i/F)^2 - \frac{1}{n}\left(\sum_i x_i/F\right)^2\right] = \frac{1}{F^2}\, s_x^2$$

If the number of words in a block is decreased to $F/2$, the number of blocks increases to $2n$. The new rate variable p_i is equal to $y_i/(F/2) = 2y_i/F$, where y_i represents the actual word counts in these smaller blocks. Calculating the variance for this new rate yields:

$$s_p^2 = \frac{1}{2n-1}\left[\sum_i (2y_i/F)^2 - \frac{1}{2n}\left(\sum_i 2y_i/F\right)^2\right] = \left(\frac{2}{F}\right)^2 s_y^2$$

The ratio of the standard deviation of p, the new rate variable in the shorter blocks, to the standard deviation of the original rate variable r is:

$$\frac{s_p}{s_r} = \left[\frac{(2/F)^2\, s_y^2}{s_x^2/F^2}\right]^{1/2} = 2\,\frac{s_y}{s_x}$$

The value 2 in this ratio represents the doubling of the number of samples when the block length is halved. Thus, when the sample length is decreased by half, the standard deviation of a word rate increases by twice the ratio of the standard deviation of the actual word counts on which the rate variables are based.

The general trend of the graphs is thus partly due to using "rate of occurrence" while decreasing the length of the samples being examined. The other factor, the change in the standard deviations of the counts themselves, will differ for each word and according to author. (Statistics for the word rates rather than those for the counts are presented in the graphs because rates are used in the analysis presented in the next chapter.) In summary, these graphs do not

provide a clear answer to the question of minimum sample size, but they do suggest that the variability begins to increase markedly when samples of less than about 1000 words are examined. Again, the analysis of small samples of known authorship is the most reliable way of determining the length of the shortest sample of text that can be assigned using a particular set of variables and a given statistical procedure. (Section 6.5.3 will describe the results of such tests for these word-rate variables and the discriminant analysis procedures developed in the next chapter.)

5.6 Summary

Distinctiveness ratios and t tests have been used to find frequent function words that might be used to distinguish text samples by Fletcher from those by Shakespeare. Some common grammatical word classes have also been examined to determine if pooled counts of their forms could be useful variables. Several forms of modal verbs (*must* and *dare*), occur at significantly different rates in the playwrights' texts, but the two authors' use of other modal verbs is not very different. Of the other groups examined, only personal pronouns showed some promise.

However, these two groups and a number of individual words were eliminated from consideration when variation within Shakespeare's texts was examined in more detail. Grouping the acts of plays by period of composition and genre showed that a number of the word and word-class variables varied significantly according to these classifications. Variables with significant internal variation were only retained if the rate of occurrence in each sub-group was significantly different from the overall Fletcher rate.

The selection process eventually resulted in a set of 14 individual word markers. In addition to these, counts of a number of less frequent words were pooled to form a "infrequent marker set" variable for each author. (Appendix E contains the counts for all 16 markers in every scene in all 34 plays examined in this study.) A number of the final set of 16 variables were shown to be significantly

correlated. An examination of the increasing variance of these variables when counted in successively shorter blocks of text provides some indication of the relationship of sample length and within-author variation.

While these 16 variables may be the most effective word-rate variables for distinguishing samples of Shakespeare and Fletcher, a successful analysis of *Henry VIII* and *The Two Noble Kinsmen* will require an appropriate statistical technique for their analysis. The problems of correlation and the form of word-frequency distributions suggest that a more sophisticated procedure should replace the univariate significance tests (such as χ^2 and the exact tests) used by Morton, Smith and Merriam. The next chapter introduces discriminant analysis procedures and evaluates the effectiveness of the combination of these procedures and these variables on texts of known authorship.

Chapter 6

Discriminant Analysis of Word Rates

This chapter introduces the statistical technique of discriminant analysis as a means of evaluating the word-rate variables isolated in the last chapter. In Chapter 3 it was observed that much of the discussion regarding the statistical validity of Morton's and Merriam's techniques centered on the use of χ^2 tests. Results presented in Chapter 4 for collocations and proportional pairs showed that the statistical independence of pairs of these tests could not be safely assumed. Multivariate procedures that allow for the correlation of literary features should be employed to analyze such features. Given the complex distributional models that often describe word occurrences or sentence length, it would also be desirable to develop methods that are at least robust from departures from distributional assumptions. True distribution-free methods would be even more satisfactory.

The general principles of discriminant analysis allow for the interrelationships between variables in the classification of disputed samples. In fact the procedures can often turn significant correlation into an advantage. In *Discrimination and Classification,* Hand illustrates how two variables that are not among the best individual discriminators can still be the most effective pair when used together [45, p. 122 and 146]. The multivariate relationships between variables can be very important but are usually too complex to be noted by human observation.

Computers come into their own in such situations, and the methods that will be described rely on machine processing. The classification procedures involve complex calculations; moreover, these must be performed repeatedly for a large number of samples to ensure the accuracy of a given *classifier,* which is a set of variables used with a particular procedure. Perhaps it is unfortunate that more simple methods do not suffice, since many non-scientists will simply view the machine and program that embodies these procedures as a "black box" that takes in data and then mysteriously and incomprehensibly produces answers. However the basic principles of discriminant analysis are intuitively appealing (especially for the nearest-neighbor methods), although the statistical justification, the mechanics of implementation and the theoretical niceties require significant mathematical skills.[1] Certainly the procedures are easier to understand than the distributional techniques developed by Mosteller and Wallace [113] and reviewed in Section 3.4.1. Indeed a number of humanities researchers have made use of statistical software packages for discrimination or cluster analysis (for example, Ledger's cluster analysis of letters in words in Greek texts [68] and Baillie's discriminant analysis of syntactic features in Shakespeare and Fletcher [5]). If university computing centers can supply satisfactory advice on statistical techniques and software then perhaps future humanities researchers will become as familiar with these multivariate techniques as their colleagues in the sciences and social sciences.

This section will first discuss the basic principles underlying discriminant analysis techniques, including the concept of a measurement space and basic statistical decision theory. Two distribution-free methods, the kernel and the nearest neighbor methods, will be described in detail. The process of feature selection, where a subset of a larger set of variables is chosen that accurately and efficiently classifies observations, is important for the success of the method.

[1] A number of introductory books on the methods assume these skills and focus on theory. Hand's book *Discrimination and Classification* [45] is perhaps the best introduction for the non-engineer or non-mathematician. I have adopted his notation in this chapter.

Several selection approaches are used with the set of word-rate variables, and the resulting classifiers are evaluated according to the misclassification rate for samples of known authorship.

6.1 Principles of Discriminant Analysis

6.1.1 The Measurement Space

The goal of discriminant analysis is to classify unknown observations by comparing them to a number of observations of known classification. To represent observations for the purpose of comparison, each one is represented by an array of numbers which correspond to a series of measurements on that observation. One can then envisage a multidimensional *measurement space* with as many dimensions as there are measurements or variables. Each observation is then represented by a point (or vector) in the measurement space. The symbol $\mathbf{x}$ is often used in formulas to represent a vector of n measurements:

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

The assumption underlying both discriminant analysis and cluster analysis is that observation vectors from the same *class* or *population* will lie close together. Dissimilarity between objects is therefore measured by the distance between the vectors that represent them.

In the current problem the observations are text samples, the classes correspond to the two authors and the measurements are the word-rate variables isolated in the preceding chapter. But the representation based on observation vectors in a measurement space is quite general. In particular, no assumptions need to be made about the nature of the measurements. This is one reason that discriminant and cluster analysis techniques have proved to be useful in many applications. However, if the measurements are different in nature one should

consider the magnitudes of the variables' values. For example, suppose the usual Euclidean distance metric

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_i (x_i - y_i)^2} \tag{6.1}$$

were used in a procedure to classify British cities as desirable places to live according to two variables, number of pubs and number of cinemas. For most cities the values of the first variable will overwhelm the values of the second, and the distance between two cities would be:

$$\begin{aligned} d(\mathbf{x},\mathbf{y}) &= \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \\ &\simeq (x_1 - y_1) \end{aligned}$$

Any information contained in the variable "number of cinemas" would not contribute much to the measure of dissimilarity. Two courses are available in such situations. The variables can be scaled or a different distance metric can be used. Hand [45] notes that most work in discriminant analysis has used the Euclidean metric, while researchers in cluster analysis have used a variety of alternative metrics to circumvent these problems.

The variables in the current problem are identical in nature; each can be expressed in terms of "number of words" or "word rate per thousand words." However, one might argue that the data values for each word should be scaled according to their standard deviation. For example, the word *the* is very frequent; the difference between Shakespeare and Fletcher's overall rates is about 9 words per thousand, compared to *too* where the difference is about 3 words per thousand. Yet the within-class variance in both authors' texts for *too* is much lower than for *the*, and for a given comparison a difference of 1 word per thousand for both variables might correspond to a highly significant deviation from the mean for *too* but not for *the*. The t tests that were used to select these words as markers of authorship allow for this by measuring the difference in means in terms of the combined variance.

In this analysis the data has not been initially standardized. Since the word-rate variables are identical in nature, it seems desirable to accept the natural

units of the data and rely on the discriminant analysis procedures to handle the greater variance of some variables. Some procedures do make use of the different variance of each of the variables, while others do not.

6.1.2 Statistical Decision Theory

Some statistical decision theory is required to understand the theoretical basis of the procedures that will be described in this chapter and used in the next chapter to classify samples of disputed authorship. The objective is to develop a procedure that determines to which class an object belongs, given its measurement vector. Assume that for each class ω_i there is an initial probability $P(\omega_i)$ that an object belongs to that class; this probability is the *prior probability*. Classification should be based on a comparison of the *posterior probabilities* $P(\omega_i \mid \mathbf{x})$: the probability of belonging to a given class given the variable values. This principle is expressed statistically using the *Bayes minimum error rule*; the observation represented by $\mathbf{x}$ is assigned to class ω_i if:

$$P(\omega_i \mid \mathbf{x}) > P(\omega_j \mid \mathbf{x}) \text{ for all } j \neq i \tag{6.2}$$

In words, the posterior probabilities are compared for each class, and the object is assigned to the class with the largest value.

Discriminant analysis procedures make these comparisons based on vectors in a measurement space, using a *decision rule* to partition this space into regions $\Omega_i, i = 1, \ldots, n$ that correspond to the n classes ω_i. If an object's vector representation $\mathbf{x}$ lies in region Ω_i, then that object is classified as belonging to class ω_i. The boundaries between the regions associated with the classes are called the *decision surfaces*. For a one-dimensional space, a decision surface is simply a point on a number line. For two measurements it is a curve, and for three it is represented by an ordinary surface. Figure 6–1 (on page 235) illustrates a two-dimensional measurement space. The variables are the rates of occurrence of *in* and *of* in acts of Shakespeare and Fletcher, and the decision surface (determined by the kernel method described later) is shown by the solid curve.

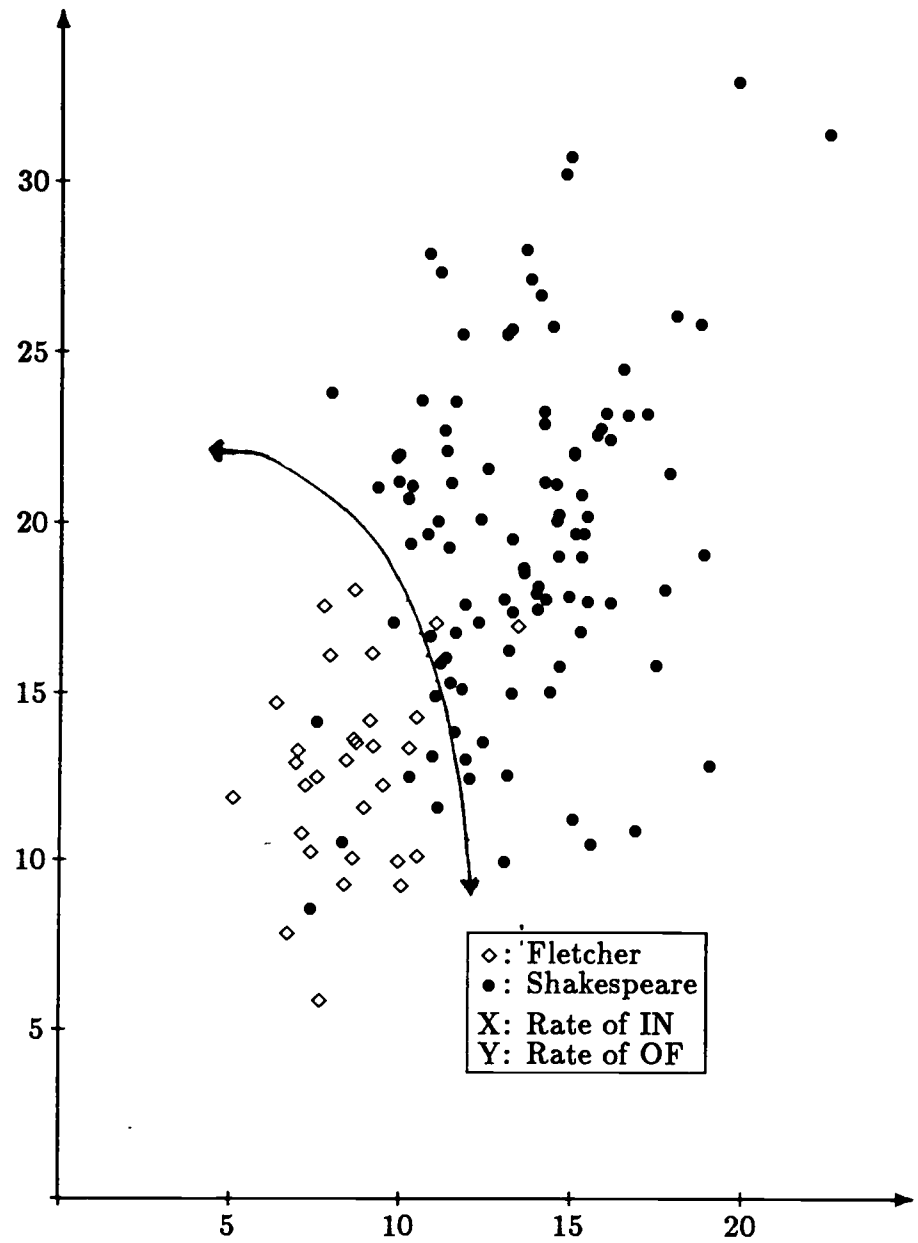


Figure 6–1: Rates of *in* vs. *of* in acts in 20 Shakespeare and 6 Fletcher plays

Sometimes the posterior probabilities $P(\omega_i \mid \mathbf{x})$ can be estimated from observations of known classification, but often Bayes' theorem is used to express (6.2) in terms of the prior probabilities $P(\omega_i)$ and the *class-conditional probability density functions* $p(\mathbf{x} \mid \omega_i)$ and the overall probability of occurrence $p(\mathbf{x})$:

$$P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i)P(\omega_i)}{p(\mathbf{x})} \tag{6.3}$$

Now the Bayes minimum error rule can be expressed as follows: allocate the observation represented by $\mathbf{x}$ to class ω_i if:

$$p(\mathbf{x} \mid \omega_i)P(\omega_i) > p(\mathbf{x} \mid \omega_j)P(\omega_j) \text{ for all } j \neq i \tag{6.4}$$

The probability density function (often abbreviated to *pdf*) describes the occurrences of $\mathbf{x}$ for each class. In many cases the form of this distribution is known, and the parameters can be estimated from the sample data.

In Equation 6.3, note that $p(\mathbf{x})$, the overall probability that $\mathbf{x}$ occurs, is the sum of the posterior probabilities or

$$\sum_{i=1}^{n} p(\mathbf{x} \mid \omega_i)P(\omega_i)$$

Since it is common to both sides of the inequality it does not appear in Rule 6.4.

In many situations (including the authorship problem at the center of this dissertation) only two classes are present, and Rule 6.4 can be more easily expressed in terms of a likelihood ratio:

$$\frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} \gtrless \frac{P(\omega_2)}{P(\omega_1)} \Longrightarrow \mathbf{x} \in \begin{cases} \Omega_1 \\ \Omega_2 \end{cases} \tag{6.5}$$

Theoretically the Bayes minimum error rule, as the name implies, represents the best decision rule for a given set of variables. In practice of course the performance of a classifier depends on how good the variables themselves are. This must be evaluated from the sample data or through theoretical analysis if the distributions are known.

Mosteller and Wallace discuss the interpretation of prior probabilities at length in *Inference and Disputed Authorship* [113, pp. 56–57]. They work with

odds (ratios of probabilities). For example, the odds that $\mathbf{x}$ belongs to ω_1 can be represented as:

$$\frac{P(\omega_1 \mid \mathbf{x})}{P(\omega_2 \mid \mathbf{x})} = \frac{P(\omega_1)\, p(\mathbf{x} \mid \omega_1)}{P(\omega_2)\, p(\mathbf{x} \mid \omega_2)} = \frac{P(\omega_1)}{P(\omega_2)} \times \frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)}$$
$$= (\text{initial odds}) \times (\text{likelihood ratio}) = \text{final odds}$$

They maintain this factorization of the odds of authorship into the product of the ratios of the probability density functions and prior probabilities throughout their study. This allows different personal evaluations of the prior probabilities to be used in conjunction with the results of their function word analyses. In their study they discover such effective discriminators that most differences in initial odds do not affect the allocation of the disputed papers.

In their discussion of initial odds [113, pp. 56–57], Mosteller and Wallace state that specifying the prior probabilities is "often considered the major obstacle to using Bayes' theorem." This concern is not echoed in Hand's book (or any other book introducing these techniques that I have read). This may reflect the increasing acceptance of the Bayesian aspect of discriminant analysis procedures since the publication of Mosteller and Wallace's book in 1964. They note that their use of Bayesian methods for estimating the parameters for the probability density functions is a "more critical use of Bayes' theorem."

In some experimental situations the prior probabilities reflect the number of observations of each class in the design set. This would be inappropriate in most studies of authorship. One could choose prior odds on the basis of a personal evaluation of the historical, stylistic and linguistic evidence. Such evaluations cannot help but be subjective, and in the case of *Henry VIII* and *The Two Noble Kinsmen* scholars' views vary tremendously. The reason for using discriminant analysis methods is to evaluate the writers' use of function words as evidence of authorship. Bayes' theorem could certainly be used to provide a means of combining function word usage with other forms of evidence, but one can argue that such evidence should not be incorporated in the decision rule based on word-rate variables.

Equal prior probabilities will be assumed in the analysis that follows. One should note that the ratio of the posterior probabilities calculated under this assumption corresponds to the likelihood ratio used by Mosteller and Wallace. Thus my results can be adjusted to reflect a personal evaluation of other evidence by using multiplication and one's opinion of the initial odds.

6.1.3 The Reject Option

One way to reduce the error rate would be to recognize borderline cases where the results of classification are doubtful. Such observations would not be assigned to any class. To implement a *reject option,* one defines a region in the measurement space where points are not classified. The highest proportion of points that are incorrectly classified will lie in the region close to the decision surface. The measurement space is thus divided into two complementary regions, the acceptance region (where classification takes place as described above) and the rejection region.

For a Bayes optimal classifier, an observation is assigned to the class with the maximum posterior probability: $\max_i P(\omega_i \mid \mathbf{x})$. The probability that $\mathbf{x}$ actually belongs to another of the classes is $1 - \max_i P(\omega_i \mid \mathbf{x})$. If this value exceeds a specified rejection threshold t then the point is not classified. A decision rule which embodies this idea is:

$$\text{if } \max_i P(\omega_i \mid \mathbf{x}) \begin{cases} > 1 - t \text{ then classify } \mathbf{x} \\ < 1 - t \text{ then reject } \mathbf{x} \end{cases} \tag{6.6}$$

Note that for smaller values of t more points are rejected and fewer classified.

Fukunaga describes how the theoretical error rate can be calculated if the relationship between the rejection threshold t and the rejection rate $r(t)$ is determined [40, pp. 154–157]. This allows one to use unclassified samples to evaluate the error rate, which may be important in some applications where the classification of samples is expensive. In a Jacobean textual problem the number of text samples is limited, but the reject option can still be used to identify samples that

may be classified incorrectly. Note that for two classes $P(\omega_1 \mid \mathbf{x}) + P(\omega_2 \mid \mathbf{x}) = 1$; thus values of $t > 0.5$ would lead to the classification of all observations. In the authorship problem being investigated here, the value of t represents the maximum value of the posterior probability for the class that is not selected. Values like .4 or .45 could be used to identify doubtful assignments.

6.2 Distribution-free Methods

As noted above, if one can determine the posterior probabilities $P(\omega_i \mid \mathbf{x})$ or the probability density functions $p(\mathbf{x} \mid \omega_i)$ then classification is straightforward. If the general form of the density functions is known then the sample data can be used to determine the parameters of these functions or a related discriminant function. The most common situation in which *parametric estimation* is satisfactory is when the distributions are multivariate normal. A great deal of attention has been focused on discrimination methods for use in situations where normality can be demonstrated or justifiably assumed.

Mosteller and Wallace demonstrated that occurrences of many function words are distributed in the writings of Hamilton and Madison according to the negative binomial distribution. The counts for all but one of the individual marker words isolated in the preceding chapter also fit this distribution. (Appendix B provides the details of this analysis of word counts.) It thus seems unlikely that the multivariate normal procedures can be used with word-rate variables in these texts. While normality is often a reasonable assumption, this should be verified using goodness-of-fit tests on the word rates themselves.

To test each variable for univariate normality, the statistical package SAS was used to perform a Kolomogorov goodness-of-fit test for the normal distribution. The values tested were the word rates per thousand in all scenes in the design set that contained at least 1000 words. Table 6–1 shows that most of the word-rate variables are not normally distributed: about half for the Fletcher scenes and over two-thirds for Shakespeare.

Marker	*rate*		arcsin $\sqrt{rate}$	
	Fl	Sh	Fl	Sh
all		•		•
are				
dare	•	•	•	•
did	•	•	•	•
in	•			
must	•	•		•
no		•		
now		•		
of				
Infreq-Sh+		•		
Infreq-Fl+				
sure	•	•		•
the		•		
these	•	•		•
too		•		•
which	•	•	•	•

Those word-rate variables marked with the symbol • fail the Kolomogorov goodness-of-fit test for the normal distribution at the 5% level of significance.

Table 6–1: Non-normal markers in scenes ≥ 1000 words

The table also shows that the arcsine transformation (arcsin $\sqrt{x}$), apparently used by Larsen and Rencher with word frequencies [66], improves the situation slightly. But one-half of the markers are still non-normal in Shakespeare. Snedecor and Cochran describe the arcsin transformation, noting that it was developed for binomial proportions [151, p. 290]. However, if the numbers of trials (in this case, total number of words in a scene) in the set of observations vary widely, then they recommend a weighted analysis in the angular scale. Since Larsen and Rencher dealt with 1000 word blocks from the *Book of Mormon,* their use of the arcsine transformation may have been justified. The question of collaboration in *Henry VIII* and *The Two Noble Kinsmen* involves scenes of varying length. Snedecor and Cochran's suggestion for a more sophisticated transformation was not employed in this case. Instead it was decided to use distribution-free methods, which do not require any information about the form

of the probability density functions. These methods have become increasingly popular in the last decade.

Distribution-free methods suffer from several disadvantages in comparison to parametric procedures. One practical disadvantage of distribution-free methods is that all sample observations must be continually accessible in order to calculate the density estimates. This contrasts with parametric estimation where only a few population parameters are used in the calculations. For this research I have not been faced with severe computing resource limitations, so this has not been a problem.

In addition, Hand presents a theorem (due to Rosenblatt) stating that any non-parametric estimate of a probability density function based on a finite numbers of samples will be biased. The meaning of *bias* in this case is that of "bias of estimation" as opposed to "bias of selection." (Thompson discusses these meanings and other details of estimation in Part III (Vol. 2, No. 2) of his *ALLC Bulletin* series on literary statistics [161].) A biased estimator is one whose value does not tend to the value of the parameter as the number of samples is increased without limit. This bias can be reduced by increasing the number of samples, but when the data comes from literary texts this is often not a practical option.

The degree of bias inherent in a pdf estimate is embodied in how well the estimate can represent irregularity in the true pdf. Distribution-free estimation methods rely on the sample points to describe the pdf. But the function best described by the sample observations is one made up of a series of probability "spikes" at each point; observations are likely to occur where the sample points actually do occur and nowhere else. Somehow the estimate must be spread out over the region between observations, and therefore each distribution-free method requires some form of *smoothing parameter*. If the estimate is oversmoothed then it may not reflect local fluctuations in the true pdf, but if it is undersmoothed then it degenerates into a collection of small regions of high probability. Determining the best value for a smoothing parameter is an important problem to be faced in distribution-free approaches.

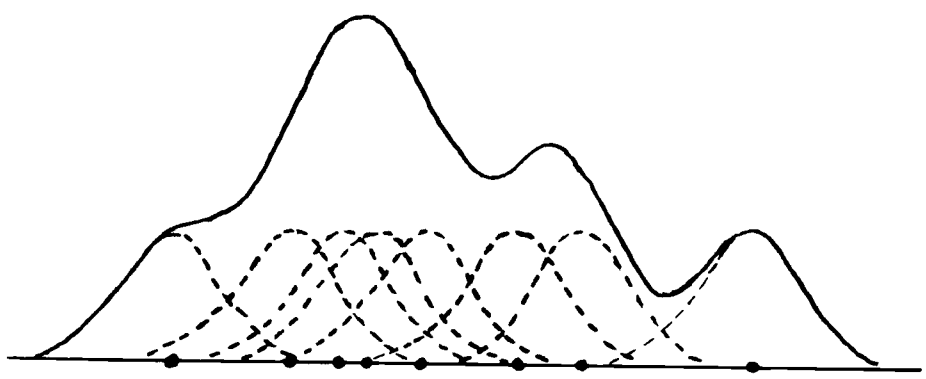
An advantage of distribution-free methods lies in their robustness; it is often difficult to predict how much effect incorrect assumptions regarding the parametric form of the density functions could have. While Fukunaga suggests that parametric methods should be used when "sufficient" knowledge of the density is available, perhaps distribution-free methods should be employed if there is any doubt. Goldstein's early study indicated that the kernel and nearest neighbor methods performed surprisingly well in comparison to their parametric counterparts [42]. This has been supported by Remme, Habbema and Hermans's detailed comparison of the effectiveness of the kernel, linear and quadratic methods with several data distributions [122]. They conclude that "the present practice of nearly exclusive use of LDA [linear discriminant analysis] cannot be justified by our results."

6.2.1 Kernel Estimators

Kernel (or Parzen) estimators provide one mechanism for estimating the value of the class-conditional probability density functions at a given point in the measurement space. Consider what information each observed sample contributes to an estimate of the class-conditional pdf at a specified point. For a point $\mathbf{x}_j \in \Omega_m$ it is clear that the pdf has non-zero value at the point itself, and if the pdf is assumed to be continuous then $p(\mathbf{x} \mid \omega_m)$ should also assume non-zero values close to $\mathbf{x}_j$. As the distance increases from this observation less information about the estimate can be inferred from this point.

When estimating the pdf at point $\mathbf{y}$ the amount of information gained by observing a point $\mathbf{x}_j \in \Omega_m$ can be represented by a *kernel* function $K(\mathbf{y} - \mathbf{x_j})$. Such a function has its maximum value at $\mathbf{x_j}$, and monotonically decreases as the distance from this point increases. The estimate of the pdf for class ω_m is based on the sum the contributions from all the observations from this class divided by their number n_m:

$$\hat{p}(\mathbf{y} \mid \omega_m) = \frac{1}{n_m} \sum_{j=1}^{n_m} K(\mathbf{y} - \mathbf{x_j}) \tag{6.7}$$



Note: Reproduced from Devijver and Kittler's *Pattern Recognition: A Statistical Approach* [26, p. 427].

Figure 6–2: Estimating a univariate pdf with normal kernels

As Hand notes, the estimator will satisfy the conditions required of a probability density function if the kernel function itself meets these conditions [45, p. 26].

The estimate of the pdf for a given class at a given point $\mathbf{x}$ depends on the kernel function's range of influence which is determined by a smoothing parameter h_m. (The smoothing parameter is sometimes referred to as the kernel function's *bandwidth*.) The value of h_m for a given estimator will certainly depend on the number of samples. A small number of points will require more smoothing, while a large number will better reflect the true continuous density function and will require less smoothing. Thus h_m should be a function of the number of observations n in the class under consideration:

$$\lim_{n \to \infty} h(n) = 0$$

Fukunaga and Hand summarize several other conditions on h_m and K which ensure that the kernel estimator is asymptotically unbiased.

A number of forms for kernel functions have been suggested, but Gaussian (normal) kernels have desirable properties and are frequently used. Figure 6–2

provides a one-dimensional illustration of how the normal kernels associated with eight observations combine to produce an estimate of the pdf. The mathematical formula for a one-dimensional normal kernel estimator is:

$$\hat{p}(\mathrm{y} \mid \omega_m) = \frac{1}{n_m h_m} \sum_{i=1}^{n_m} (2\pi)^{-1/2} \exp\left[-\frac{1}{2}\left(\frac{y - x_i}{h_m}\right)^2\right]$$

Varying the value of the smoothing parameter h_m would make the individual bell-shaped curves either taller or flatter, making the pdf estimate either spikier or smoother.

The corresponding general form of a multivariate normal kernel function is:

$$\begin{aligned} \hat{p}(\mathbf{y} \mid \omega_m) &= \frac{1}{n_m h_m^d} \sum_{i=1}^{n_m} (2\pi)^{-d/2} \exp\left[-\frac{1}{2} \frac{(\mathbf{y} - \mathbf{x_i})'(\mathbf{y} - \mathbf{x_i})}{(h_m)^2}\right] \\ &= \frac{1}{n_m h_m^d} \sum_{i=1}^{n_m} (2\pi)^{-d/2} \exp\left[-\frac{1}{2} \sum_{j=1}^{d} \left(\frac{y_j - x_{ij}}{h_m}\right)^2\right] \end{aligned} \tag{6.8}$$

In this equation d is the number of dimensions; $\mathbf{y}$ is the point at which the estimate is being made; and the vectors for the design-set observations for class ω_m are represented by $\mathbf{x}_i$, $i = 1, \ldots, n_m$. Since the value of h_m is the same for each dimension, the shape of the kernel function will be the same in each dimension (that is, radially symmetric). If the variances for each variable are different enough then it might be desirable for the kernel to allow for these differences. In any case, some estimation method must be used to select the values of the smoothing parameter for each class.

Silverman describes a number of methods that can be used to determine the best value for a smoothing parameter [142]. One automatic method is the maximum likelihood method outlined by Habbema, Hermans and van den Broek as part of the program described in "A Stepwise Discriminant Analysis Program Using Density Estimation" [44]. Maximum likelihood estimation is a general statistical technique for choosing a value for an estimate. (Again, Thompson provides a useful introduction to this method in Part III of his *ALLC Bulletin* series on literary statistics [161].) In this approach the data observations are viewed as a series of values that are a function of the parameter, now regarded

as a variable. The best value to use for the parameter is that which maximizes this function; in other words, choose the value which is most likely to have given rise to the actual observed data points.

To use this estimation method for the smoothing parameter, one must maximize a likelihood function product of the pdf estimates at the points in the design set. This process produces an optimum value $\hat{h}_m = 0$, which results in a pdf estimate composed of a probability spike at each observation and zero density elsewhere. The procedure suggested by Habbema *et al.* circumvents this by employing a *leaving-one-out* modification of the maximum likelihood method. In calculating the pdf at each point for the likelihood function, that point itself is left out, thus avoiding the useless result.

C. G. G. Aitken of the University of Edinburgh's Statistics Department has developed a program called KERCON modeled on the procedure described by Habbema, Hermans and van den Broek. KERCON designs a classifier using a multivariate normal kernel estimator, allowing the kernel shape to vary in each dimension according to the variance of the variable corresponding to that dimension. To accomplish this the algorithm follows Fukunaga's suggestion (most clearly explained by Silverman [142, pp. 77–78]). First the data values are standardized by dividing by the standard deviation. The calculation of a single smoothing parameter for all dimensions (using the maximum likelihood method outlined above) is then based on this transformed data. Finally the variables are transformed back to their original units and the smoothing parameter is introduced into the normal kernel formula. This is equivalent to using the following estimate of the density function:

$$\hat{p}(\mathbf{y} \mid \omega_m) = \frac{1}{n_m \hat{h}_m^d s_{m1} \ldots s_{md} (2\pi)^{d/2}} \sum_{i=1}^{n_m} \exp\left[-\frac{1}{2}\sum_{j=1}^{d}\left(\frac{y_j - x_{ij}}{\hat{h}_m s_{mj}}\right)^2\right] \qquad (6.9)$$

where s_{mj} is the sample standard deviation of the data points of class ω_m for variable j. Comparison of this formula with (6.8) indicates that in effect the value $\hat{h}_m s_{mj}$ is being used to weight the contribution to the pdf estimate for

dimension j, but only one smoothing parameter h_m must be calculated for each class.

Written in FORTRAN77, program KERCON runs under the university's EMAS operating system and on a Gould Powernode 9800 machine running a version of Berkeley Unix. The process of maximizing the likelihood function to estimate the smoothing parameters requires the calculation of a derivative, and numerical methods to solve the resulting equations are necessary. KERCON makes use of routine E04ABF of the NAG FORTRAN subroutine library to perform the maximum likelihood estimates.[2] KERCON does not reflect recent advances in kernel estimation described in Silverman's recent monograph [142]. For example, because densely concentrated regions should require less smoothing than sparsely concentrated ones, it would be desirable to allow the smoothing parameter to take on different values in different regions of the measurement space. Silverman describes how a *variable kernel* can accomplish this. Program KERCON uses a *fixed kernel,* which may give rise to spurious noise in the pdf estimates in the distribution tails. Another recent development is a *least squares* method for automatically determining the smoothing parameter. This appears to be slightly superior to the maximum likelihood method. Although program KERCON does not reflect the latest developments in this rapidly advancing area of statistics, it plays a central role in the discriminant analysis described in later sections.

6.2.2 Nearest neighbor methods

The general principle behind a kernel estimator, illustrated by Equation 6.7, is simple enough. On the other hand, the statistical calculations required to implement a multivariate normal kernel function (as described by Equation 6.9) are certainly not straightforward for the non-statistician. *Nearest neighbor* methods

---

[2]The NAG (Numerical Algorithms Group) library is a collection of subroutines for numerical calculations. For complete information about the library contact: NAG Central Office, Mayfield House, 256 Banbury Road, Oxford, United Kingdom.

provide a more simple and intuitively attractive approach to distribution-free classification.

In one approach, the nearest neighbor method can be used to estimate the class-conditional pdf at a point $\mathbf{x}$. Consider a region L centered at $\mathbf{x}$ which contains a fixed number of points k from a chosen class ω_m. For the two-dimensional example shown in Figure 6–1, such a region would be the circle centered at the specified point and just including k observations from one of the classes. The area of this circle is determined by the distance from $\mathbf{x}$ to the k-th nearest neighbor. In regions of high probability density such a circle would be expected to have a small area and in sparsely populated regions it would have a larger area. Generalizing for a multidimensional space, the volume of the hypersphere centered at point $\mathbf{x}$ and occupied by k points is related to the probability density function at $\mathbf{x}$. If the volume of region L is V, and θ is the probability that a point will fall in region L, then an estimate of the pdf is:

$$\hat{p}(\mathbf{x} \mid \omega_m) = \theta/V$$

The estimate θ/V is the average value of $\hat{p}(\mathbf{x} \mid \omega_m)$ in the region L and will be a better estimate for small L. While the probability θ is a function of the pdf, it can be estimated from the proportion of points that actually lie in L, which is k/n_m. This results in the *k-NN estimator:*

$$\hat{p}(\mathbf{x} \mid \omega_m) = \frac{k}{n_m V} \tag{6.10}$$

The smoothing parameter for nearest neighbor procedures is the number of points k.

Hand shows that one can arrive at the kernel method by fixing the volume in Equation 6.10 and determining the number of points k that this includes [45, p. 31]. On the other hand, nearest neighbor procedures fix k and allow V to vary according to the value of the probability density function in different regions of the measurement space. This avoids one drawback of the fixed kernel method, in which the smoothing parameter h is constant throughout the space. Theoretical results show that k should increase with larger values of n_m and that the volume

V should decrease. Goldstein's results confirm that, for a given number of design-set observations n_m, a suitable choice for k is $\sqrt{n_m}$ [42]. Fukunaga notes a disadvantage of this estimator; the volume V (hence k) should be kept small to obtain a relatively uniform density function in the hypersphere, but this results in a decrease in the accuracy of the estimator unless n_m is large [40, p. 178]. Hand describes a theoretical disadvantage of the k-NN estimator. It is not a true probability density function; when integrated over the entire measurement space the result is infinity instead of one. This may not be a practical disadvantage in some situations [45, pp. 32 and 43].

Nearest neighbor methods can result in a procedure for classification that is more simple and direct than estimating the density function (and less troubled by the disadvantages noted above). Instead of using only the points from a single class, region L of volume V is determined by choosing the nearest k points from any class. Of these k points, k_m will belong to class ω_m. Since this region is defined differently than above, the estimator $\hat{p}(\mathbf{x} \mid \omega_m) = k_m/(n_m V)$ for each class is slightly different from that defined by Equation 6.10. An estimate of the overall probability of the occurrence of $\mathbf{x}$ in region L is the proportion of points that fall in the region, divided by the volume:

$$\hat{p}(\mathbf{x}) = \frac{k}{nV}$$

where n is the total number of points, $\sum n_m$. According to the Bayes minimum error rule (6.2, page 234) $\mathbf{x} \in \Omega_i$ if

$$\begin{aligned} P(\omega_i \mid \mathbf{x}) &= \max_j P(\omega_j \mid \mathbf{x}) = \max_j \frac{\hat{p}(\mathbf{x} \mid \omega_j) P(\omega_j)}{p(\mathbf{x})} \\ &= \max_j \frac{(k_j/n_j V) P(\omega_j)}{k/nV} = \max_j \frac{k_j}{k} \frac{n}{n_j} P(\omega_j) \end{aligned} \tag{6.11}$$

Since k and n are the same for each class, this is equivalent to maximizing $(k_j/n_j)P(\omega_j)$. This is known as the *k-nearest neighbor* (k-NN) *classification rule*; it is certainly much simpler to implement than most kernel estimators.

If each class has equal prior probabilities and the same number of samples in the design set, then $\mathbf{x}$ is simply classified as belonging to the class with the

largest number of nearest neighbors. (In this situation k is usually chosen to be an odd number in order to avoid the possibility of having the same number of nearest neighbors from each class.) In particular, if $k = 1$ then the *nearest neighbor* (NN) classification rule assigns $\mathbf{x}$ to the same class as the observation closest to it in the measurement space. While this would appear to make very little use of the information available from the design-set data, theoretical results show that the NN rule has an asymptotic error rate bounded by twice the Bayes minimum error rate. This rather remarkable result has led to widespread interest in the nearest neighbor rule (especially in the field of pattern recognition), since it implies that half the classification information in the measurement space is contained in the nearest neighbor, even if the number of observations is infinite. (This result is due to Cover and Hart [24]; the theory regarding the error bounds for the NN and k-NN rules is discussed thoroughly by Devijver and Kittler [26].)

These results are based on the assumption that each class is composed of equal numbers of sample observations. It is interesting to consider carefully the use of the k-NN classification rule (6.11) with two classes having unequal numbers of observations or unequal prior probabilities. For example, the NN rule will assign an observation to the same class as its nearest neighbor no matter what the values of $P(\omega_i)$ or n_i. Although k-NN techniques are popular in the area of pattern recognition, the possibility of different classes having unequal numbers of observations or prior probabilities is not often discussed in the literature. (The development given above follows Hand's, a statistician.) Fukunaga assumes that the proportion of observations in each class reflects the prior probabilities; the two factors thus cancel out [40, p. 179]. Devijver and Kittler do not consider either factor at all in their description of the technique.

An 1979 article by Brown and Koplowitz [21] proposes that weighted distances be used with the NN rule when class ratios do not approximate the prior probabilities. If $\mathbf{x} \in \omega_i$ then the normal Euclidean distance metric is multiplied by $(n_i/nP(\omega_i))^{1/d}$, where d is the number of dimensions. Although they discuss the asymptotic performance for 1-NN, they do not extend their analysis to larger

values of k. Using a weighted distance to penalize neighbors from classes having a higher proportion of observations is intuitively attractive. However, it is unclear if this method is compatible with the reasoning used to derive (6.11), where a comparison is made of the density function estimates in a common region.

A major practical disadvantage of k-nearest neighbor classification is that there is no statistically justified method for determining what value of k to use. Hand can only suggest trial and error, choosing a value that produces the lowest number of misclassifications on samples of known classification. Such an approach is unsatisfactory in a process such as feature selection (described below in Section 6.4). To determine the best subset of variables, selection procedures usually design and compare many different classifiers. Often each comparison requires substantial time and resources. If a number of values of k must be tested for each set of variables, then the potentially large number of comparisons increases substantially.

Nearest-neighbor classification rules are easy to calculate once the distances between all combinations of points in the design space have been calculated. In addition, the statistical package SAS (available on a large number of computer systems) includes a procedure NEIGHBOR that classifies observations according to the k-NN classification rules. Unfortunately, the SAS procedure does not allow for different numbers in the classes (unless the proportions n_m/n reflect the prior probabilities $P(\omega_m)$). Instead of assigning to the class corresponding to $\max\{(k_j/n_j)P(\omega_j)\}$ NEIGHBOR uses $\max\{k_jP(\omega_j)\}$. To circumvent this feature one must supply values for prior probabilities which reflect the different number of samples in each class. These values (call them P_j') must be proportional to $P(\omega_j)/n_m$ but must also add up to 1 (since the software checks for valid input values for the prior probabilities). For two classes and equal prior probabilities, algebra shows that the required values are $P_1' = n_2/n$ and $P_2' = n_1/n$.

6.2.3 Examples Using the Kernel and k-NN Methods

As an example, consider the measurement space based on *in* and *of* illustrated in Figure 6–1 (on page 235). These frequent prepositions are not the best pair of variables for distinguishing acts of Shakespeare and Fletcher and are only used to illustrate the principles of discriminant analysis. First, high rates of *in* appear to correspond to high rates of *of* in the Shakespeare acts. Using SAS to evaluate Kendall's τ correlation coefficient shows that the rates for the 101 Shakespeare acts are positively correlated ($\tau = .16$, $p = 1.9\%$). The correlation coefficient for the 30 Fletcher scenes is not significant for the illustrated data ($\tau = .15$, $p = 23\%$).[3]

The non-linear decision surface shown in Figure 6–1 is based on a normal kernel estimator as determined by program KERCON. By repeatedly creating a large number of data points and classifying them, points where the posterior probabilities were equal were isolated. The classes are not as well separated as one would like, but (as indicated by the curve) the kernel method misclassifies 10 of the 131 total acts (7.6%). (The misclassification figures cited in this section are determined using the leaving-one-out method, described in the next section.) Two of the misclassified observations are close to the decision surface: Shakespeare's *TGV* Act 3 where $P(\text{Fl} \mid \mathbf{x}) = 0.55$, and Fletcher's *Deme* Act 1 where $P(\text{Sh} \mid \mathbf{x}) = 0.58$. A number of Shakespeare acts are deep in Fletcher territory; for three of these the posterior probabilities $P(\text{Fl} \mid \mathbf{x})$ are over 90%. The smoothing parameters (calculated using the maximum likelihood technique) are $h_{\text{Sh}} = 0.50$ and $h_{\text{Fl}} = 0.76$. As expected, the smaller number of Fletcher observations required more smoothing.

Table 6–2 shows the results of k-NN neighbor classification for a number of values of k. The misclassification rates (again using the leaving-one-out method)

[3]However, when these two variables are measured in scenes of 1000 words or more, the correlation coefficients for both writers are significant: for 168 Shakespeare scenes $\tau = .19$, $p = .0003$ and for 54 Fletcher scenes $\tau = .25$, $p = .007$.

	Number of misclassifications			
k	Fl	Sh	Overall	
1	9	9	18	13.7%
2	3	11	14	10.7%
3	2	14	16	12.2%
4	2	16	18	13.7%
5	2	6	8	6.1%
6	2	7	9	6.9%
7	2	11	13	9.9%
8	2	12	14	10.7%
9	2	8	10	7.6%
NN with weighted distance metric:				
1	3	17	20	15.3%

Table 6–2: k-NN misclassifications for Figure 6–1

vary; the best performance is for $k = 5$ when 8 acts (6.1%) are incorrectly allocated. The nearest neighbor classifier shares the title for worst performance (with $k = 4$), producing 18 misclassifications (13.7%). The decrease in the rate of error for $k = 5$ and the increase for $k = 7$ are striking. Clearly the choice of k is quite important.

For this design set composed of 30 acts of Fletcher and 101 acts of Shakespeare, for $k = 4$ an unknown observation is assigned to Fletcher unless all four neighbors are Shakespeare acts. (If $k_{\text{Sh}} = 3$ then $k_{\text{Sh}}/n_{\text{Sh}} = 3/101 = 0.030$ and $k_{\text{Fl}}/n_{\text{Fl}} = 1/30 = 0.033$.) While one can accept the mathematics underlying (6.11) the discrete behavior of the classification rule is perhaps an unwelcome characteristic of the method. Perhaps the difference in the number of misclassifications for $k = 4$ and $k = 5$ is just due to this discrete behavior. The result using Brown and Koplowitz's weighted NN rule is also given in Table 6–2. For *in* and *of* counted in acts, their method results in 20 misclassifications for the design set (15.3%). This does not compare favorably to the unweighted rate, although

the majority of the misclassifications are now acts of Shakespeare (which might be expected since over three-quarters of the design-set observations are his).

Both distribution-free methods outperform the parametric classifiers that assume normality. For these comparisons SAS's procedure DISCRIM was used to classify the design-set observations using both the linear and quadratic discriminant functions. If the variances for the two classes are equal, the linear method is optimum; otherwise the quadratic method should be used. For the *in* and *of* data the linear function misclassifies 2 Fletcher acts and 14 of Shakespeare's, an overall rate of 12.2%. The quadratic method performs slightly better, misclassifying two fewer Shakespeare acts (9.9%). These results are slightly inferior to those achieved using the kernel method and the k-NN method (for some values of k). This supports the contention that distribution-free methods are preferable to these two parametric methods in the analysis of word-rate data.

It is interesting to return to the collocations and proportional pairs that were examined in Section 4.5 on page 175. There the results of a univariate significance test for six variables were multiplied to produce a likelihood ratio for the test-set samples. Four of these thirty acts were incorrectly assigned by this process. The multivariate kernel method also misclassifies 3 of these same acts. In addition, its results differ from the previous method for two acts; it correctly assigns Act III of *Valentinian* but misclassifies Act I of *The Tempest.* Thus, a multivariate method can lead to a result that differs from that produced by a combination of results from a number of univariate tests.

The results obtained when reclassifying the design set are not unimpressive. Four of the 131 acts are misclassified, an error rate of 3.1%. These features can be used to distinguish between large samples by these two authors, but as noted earlier they do not occur frequently enough to be useful in the examination of scenes. As will be seen, the rates of the marker words discovered in Chapter 5 can be used with greater success.

6.3 Assessing a Classifier's Performance

An important component of a discriminant analysis application is the evaluation of how well a classifier (the data for a set of variables combined with a particular method of implementing a decision rule) will perform on observations of unknown classification. Naturally this will depend on the variables used and the form of the probability density functions. If the probability density functions are multivariate normal then it is possible to determine the theoretical error rates from the data means, the covariance matrix and the prior probabilities for each population. Lachenbruch outlines the principles and demonstrates the calculations for this situation [65].

On the other hand procedures have been developed that estimate the error rate without making assumptions about the forms of the distributions. These are generally based on the study of misclassification rates of samples of known classification. One source of such samples is the design set itself. The misclassification rate obtained by resubstituting each observation in the design set back into the classifier based on these observations is known as the *apparent error rate*. This rate is usually an optimistic estimate of how well the classifier will perform on new data. (Note that for the nearest-neighbor method, resubstitution will correctly classify every observation, since a point will always be its own nearest neighbor.) Since the classifier reflects the characteristics of the design set, the extent to which the apparent error rate differs from the true error rate depends on how well the design set represents the population distribution. The larger the design set, the closer the apparent error rate will approximate the true rate.

One way to avoid the biased estimate produced by resubstitution is the *sample partition* (or *holdout*) method. A set of observations of known classification is not included in the design set but is reserved for use in testing the classifier. This group is known as the *test set*. A criticism of this approach is that a better classifier should result if these observations were included in the design set. Thus

the sample partition method may not make best use of the available data. Some have divided the data into design and test sets for development and testing and then recombined the sets to design the classifier for the final application. However, the testing process then over-estimates the error rate of the final classifier [20, p. 355].

Another problem with the sample partitioning approach is determining how best to divide the observations into two sets. Fukunaga shows that the numbers of observations to include in each set depends on the dimensionality of the measurement space, observing that "in many cases" more samples should be used for testing than design if the goal is to obtain an accurate estimation of the error rate [40, p. 153]. One should also note that a single test set only provides a single estimate of the true error rate; ideally a number of such sets would be used.

A method that uses the design-set observations in a more satisfactory manner is the *leaving-one-out* method. Each observation in the design set is assigned to a class using the classifier designed with that observation omitted. The proportion of observations misclassified is an almost unbiased estimate of the expected actual error rate, as shown by Lachenbruch [65]. Fukunaga notes that the value indicated by the leaving-one-out method (or the sample partition method) provides an upper bound for the true error rate. The resubstitution method provides a lower bound, but this is usually of less interest in evaluating a classifier [40, p. 149].

One possible problem regarding the leaving-one-out procedure is that n_m distinct classifiers must be designed, and for some methods of discriminant analysis the computation required may be prohibitive. However, for the two distribution-free methods outlined earlier this should not be a problem. Both the normal kernel and k-nearest neighbor methods require that the distances between every combination of points be computed and stored. To implement the leaving-one-out method for the k-NN classification procedure, the distance between the current observation and itself (zero) is ignored when counting nearest neighbors. For pdf estimation using normal kernels, Fukunaga shows that, once all pairs of

distances are calculated, the leaving-one-out method requires the same amount of computation as the resubstitution method [40, p. 176].

Since a single point is effectively being removed from the design set, one must subtract 1 from the total number of design-set points n and from the number of points n_m for its class ω_m. The NEIGHBOR procedure in SAS does not make this adjustment when reclassifying the design set, although it does ignore the point being reclassified when finding nearest neighbors. Partly because of this (and partly to check some peculiar SAS results) a program KNN was developed to carry out k-NN classification. (Written in standard FORTRAN77, this program runs on a large number of machines.)

For the kernel method, Fukunaga assumes that there is no cost in re-estimating the values of the smoothing parameter h for each of the n_m classifiers. On the EMAS version of the program KERCON, the estimation of the smoothing parameters makes up a large proportion of the processing time required for a given run. To avoid prohibitively excessive computing costs, the program calculates the smoothing parameters once using all the design-set observations in a given class. These parameters are used for the classifiers designed to allocate each observation using the leaving-one-out method.

As described at the beginning of Chapter 2 the plays used in this study were divided into a control set and a test set. The plays in the test set (Shakespeare's *Richard III, The Tempest, As You Like It* and *Anthony and Cleopatra* and Fletcher's *Monsieur Thomas* and *Valentinian*) were used to evaluate collocations and proportional pairs in Section 4.5 by comparing each test-set sample to the overall counts of the features in the control set. In using discriminant analysis to evaluate word-rate variables, the control set of 20 Shakespeare plays and 6 Fletcher plays will be used as the design set. From the previous discussion it might seem advisable to also use the test-set plays in designing a classifier; however, three of the texts used have textual complications. The version of *Richard III* used in this study is a combination of the Folio and Quarto editions, and the relationship between these editions is not absolutely clear. The two

Fletcher test texts, *Monsieur Thomas* and *Valentinian*, are not from the most authoritative editions.

For these reasons I decided to retain these six plays as a test set for discriminant analysis. The error rates of the classifier(s) used to test the scenes of *Henry VIII* and *The Two Noble Kinsmen* will be estimated using the leaving-one-out method with the design set and the hold-out method with the test set. The textual peculiarities of the three plays that make them unsuitable for use in the design set should make them more interesting tests for the assignment procedures. The corruptions that may have been introduced in these texts could certainly be present in any Jacobean text (although to a lesser degree in the design-set texts).

6.4 Feature Selection

In a discriminant analysis problem there are a number of reasons for isolating an effective subset of features from the total set of variables. Since humans classify objects by evaluating the few most important features that distinguish one class from another, a reduction of the number of variables is intuitively appealing (but perhaps less so in a stylometric authorship study, since the analysis should be based on differences in features that are not apparent to a writer). Thus one role of a feature selection process might be to determine whether a small number of variables discriminate as well as the complete set. In other applications considerations of cost are important; it may be expensive to obtain numerous measurements for a design set, or computer processing of existing data may prove too expensive or too slow. In any situation there is no point in processing variables that do not contribute to the accuracy of the classifier.

However, even in situations where each variable has been selected for its discriminating power, a reduction in the dimensionality of the measurement space is desirable. Early pattern recognition researchers discovered that, as more measurements were made on a set of observations, the misclassification rate at first

decreased but then began to climb. Since each additional variable should add information (or at least never take any away) this is counter-intuitive. Devijver and Kittler [28, pp. 187–194] and Hand [45, pp. 121–123] provide an informal discussion of the proposed theoretical explanations of this phenomenon.

Statistical theory shows that the introduction of another measurement will decrease the error rate for an infinite set of observations. However, for a finite number of design-set observations, the number of parameters defining the decision surface increases with each new measurement. Since an estimation error is associated with each parameter, eventually the cumulative effect of these errors results in a deterioration of performance on an independent set of data (although performance may continue to improve on the design set). In addition, as the number of dimensions d increases, the design-set observations become more and more sparsely distributed over the measurement space and less representative of the true density function. Thus the classifier does not generalize well for test-set observations and the true error rate increases. This can be avoided by increasing the number of observations in the design set, but this option is not open in an authorship study when the number of texts are limited.

Two methods are used to counter this effect, known as Bellman's *curse of dimensionality.* To reduce the dimensionality of a classifier, feature *transformation* (or *extraction*) can be used to map the n-dimensional observation vectors onto a feature space of fewer dimensions. An advantage of this approach is that it makes use of all the data; however, the transformation process itself may be affected by estimation errors or assume distributional forms of the density functions [28, p. 194]. (Transformation is the principle behind *canonical variate analysis,* a discriminant analysis technique which is implemented in a number of statistical software packages.)

The second method is to use a subset of variables in a reduced feature space, without transformation. Selecting the most effective subset is a substantial task; for even a small number of variables, the number of possible subsets is quite large. Several approaches to *feature selection* have been suggested; all involve

comparisons of error rates or of some criterion related to the error rate, such as class-separability measures. Since a large number of variable sets must be evaluated and compared, it is desirable that such criteria be easy to compute. However, most (if not all) such criteria require some knowledge of the forms of the density functions and are thus not suitable for distribution-free classifiers. Generally, in this situation comparisons must be based on misclassification rates in the design set or a test set.

6.4.1 Search Methods

Since classifying observations can require significant amounts of computer time, evaluating all possible subsets is often not feasible. For example, choosing the best set of 5 variables from the 16 markers found in the last chapter would involve testing

$$\binom{16}{5} = \frac{16!}{5!\,(16-5)!} = 4,368$$

combinations. Since the kernel method program KERCON requires about 90 seconds of cpu time to evaluate 5 variables in 371 scenes, exhaustive evaluation of these sets would require over 4.6 CPU days. There is no reason to think that a 5 variable set might be optimal. Certain accelerated search methods (for example, a branch and bound algorithm) have been proposed for use with separability measures (described by Hand [45]) but do not appear to be feasible for use when classifiers are judged by misclassification rates. Methods that base feature selection on misclassification counts usually make use of suboptimal search methods. These require fewer comparisons but cannot guarantee to find the best subset of variables.

The most common method employed is *forward sequential selection*, in which one variable at a time is added to the set already chosen. This method is used by Habbema *et al.* in their discriminant analysis software [44,47]. To begin the process one selects the best individual variable. Each of the remaining variables is then tested in conjunction with this initial variable, and the best pair is determined. Each of the remaining variables is tested with this pair, and the process

continues until some stopping criterion is reached. If expense is important then the process may be halted when the addition of another variable does not significantly increase the accuracy of the classifier. In any case, selection should cease when the addition of another variable decreases the accuracy of the classifier (due to of the curse of dimensionality). As noted by McKay and Campbell "dips" in the rates of misclassifications may cause premature termination of the selection process, and they suggest that any subset be compared against the full set of variables [81]. Several disadvantages of the forward selection procedure are clear. Once a variable has been chosen it cannot be removed, even if variables selected later make it redundant. In addition no account is taken of the interrelationships between the variables that have not been selected.

A strategy that partially meets these problems is *sequential backward elimination.* Beginning with a set of N variables (perhaps the entire set), the variable that decreases the classifier's accuracy least is removed to yield a set of size $N - 1$. While this is computationally more expensive than forward selection, it can provide a measure of a subset's performance against the total set. Also, combinations of variables that discriminate better together than individually are likely to be retained. A combination of the two methods to incorporate limited backtracking results in a "Plus l-Take Away r" procedure. For example, if $l = 2$ and $r = 1$ then every combination of variable pairs is tried with the existing subset of N features; after the best set of $N + 2$ is determined backward elimination is employed to choose the best subset of size $N + 1$. While this requires fewer comparisons than an exhaustive search, the numbers are still quite large; the method is not practical if comparing subsets requires significant computing time. Devijver and Kittler remind readers that suboptimal search methods are not guaranteed to find the best feature subset; even the more sophisticated approaches may not produce a better set than simpler methods for a given set of data.

6.4.2 Application to Design-Set Data

Distribution-free methods have been shown to be more stable at higher dimensions than parametric classifiers (such as the commonly-used linear discriminant function) [45, pp. 124–126]. However, for the reasons discussed at the beginning of this section, feature selection is still desirable. The kernel method was used with both forward selection and backward elimination search methods to find feature subsets that effectively classify scenes in the design set. The k-NN classification procedures were not used in feature selection since (as noted earlier) there is no justified method other than trial and error for choosing a value of k. Comparing two sets of features for several values of k would be both difficult and computationally expensive, but both distribution-free methods will be evaluated with the subset of features chosen using the kernel method.

There are practical considerations regarding kernel method estimation that make high-dimensional spaces undesirable. As noted earlier, as the the number of dimensions increases, the design-set observations become more sparsely distributed in the measurement space. Therefore the density estimates at a given point become smaller (since the integral of a density function over the entire feature space is 1). Underflow in the computer calculations eventually becomes a problem. Program KERCON's maximum likelihood estimation of the smoothing parameters also fails to cope with some large subsets of variables. This occurs when the estimate tends toward zero, which would result in no smoothing. The software stops this process at a specified lower bound, but classification of the test set in such situations shows that the classifier is quite unsuccessful for a independent set of observations.

If variable subsets are to be selected on the basis of smallest misclassification rates, small differences in the number of errors are difficult to interpret. The proportion of misclassifications is a point estimate $\hat{E}$ of the error rate, but it more realistic to consider a confidence interval around this value. Devijver and Kittler [26] show that (given some simplifying assumptions) the sample-based

estimate of the standard deviation of this statistic is

$$\sqrt{\frac{\hat{E}(1-\hat{E})}{N}}$$

where N is the total number of points classified. Using this estimate in the usual way, the 95% confidence interval is

$$\hat{E} \pm 1.96\sqrt{\frac{\hat{E}(1-\hat{E})}{N}}$$

For example, if one subset of variables produced resulted in 4 misclassifications and another 8 in 222 observations, the 95% confidence intervals are (0.1, 7.9) and (2.6, 13.4) respectively, and it is not clear that the first set is really superior to the second. However, the selection of variables in this study was based on the number of misclassifications. While recognizing the uncertainty associated with misclassification counts, a choice between two sets of variables has to be based on some criterion. While this approach is not completely satisfactory, other studies (and software products) also select variables in this manner.

A basis for resolving ties involves making another difficult choice. Frequent occurrence is a desirable characteristic for authorship markers and this factor was initially used when two sets produced the same number of errors. Forward selection was the first method used. Since previous studies by Larsen and Rencher and Mosteller and Wallace have used discriminant analysis on samples of 1000 words, feature selection was initially based on the 222 scenes from the 26 plays of the design set that were at least this size. This selection process resulted in an 8 variable subset (in order of selection) made up of

sure all must in of these now too (Set FS1)

Using the leaving-one-out method to classify the design-set samples, this classifier allocated each of the 222 scenes correctly. (The code "FS1" signifies that this is the first group chosen using forward selection. Some simple statistics for all the word variables are given in Table 5–12 on page 215. In addition, the members of each feature subset that will be described are listed in Table 6–3 on page 264.)

The univariate t' statistic can also be used to decide between two feature subsets that produce the same number of design-set misclassifications: in case of a tie, one selects the variable with the highest t' value. (As noted in the previous chapter, the probability associated with each t' value should be used as the criterion, but for large numbers of degrees of freedom t' is approximately a standard normal deviate. The number of degrees of freedom for each of the 16 marker words tested here is greater than 100.) Applying this rule, the following 8 variable subset is selected:

sure too all dare of in which the (Set FS2)

Although in this set *too* is selected much earlier than in set FS1 and *must* is ignored, the results for scenes of 1000 words or more are about the same. Set FS2 classifies all but one scene correctly.

Testing was begun using the 1000 word minimum size because samples of this length had been used in other studies. However, since a large number of the disputed scenes are smaller than this size, this choice is somewhat artificial. Both sets are not as successful when used with a design set composed of scenes of 500 words or more. Set FS1 misclassifies 8 of 371 scenes (2.2%); set FS2 misclassifies 9 (2.4%).

While these rates might be considered acceptable, it was observed that another set of words performed better than sets FS1 and FS2 on smaller samples. This group consisted of frequent words with large t' values and was chosen subjectively when I initially tested program KERCON. This result indicated that the forward selection procedure might not be choosing the best subset, and I decided to try the backward elimination approach, using the 371 scenes with at least 500 words as the design set. Ideally one should start with all 16 variables, but program KERCON often fails to estimate a smoothing parameter or encounters underflow problems for 10 or 11 dimensions. In choosing a subset of variables with which to begin the elimination process, I was again faced with basing decisions on one of the two criteria, frequency or univariate discrimination (as measured by the t' statistic). Two initial subsets were chosen; the first based

	FS1	FS2	T1	T2	FR1	FR2
all	•	•	•		•	•
are					•	
dare		•	•	•		
did						
in	•	•	•	•	•	•
must	•				•	
no						•
now	•		•		•	•
of	•	•	•	•	•	•
Infreq-Fl+				•		•
Infreq-Sh+				•		•
sure	•	•				
the		•	•	•	•	•
these	•					
too	•	•	•	•	•	
which		•	•			

Table 6–3: The words in selected subsets of features

on the most frequent words and the second on the best individual discriminators. When eliminating variables from these initial sets, tied comparisons were decided according to the criterion by which the initial set was chosen. While two possibly very different subsets might emerge from this process, each can be evaluated using the test-set observations. (One might also hope that both subsets would produce the same results on the disputed samples.)

The ten most frequent words under consideration are (in order of decreasing frequency):

the of in all no are now too must these

After eliminating *no* and *these* the misclassification rate declined or remained level. The resulting set FR1 ("FR" for "frequent") produced only 1 misallocation out of 371 scenes (0.27%), a better performance than the sets chosen by forward selection.

The second selection process, based on words chosen by t' value, eliminated *are* and *sure* from the following starting set (ranked by decreasing value of t'):

in of too the sure all which dare now are

Note that *sure* was the first word picked using forward selection. Since it was the first variable eliminated from this set, the information it contributes to the classifier is redundant, although *sure* is the best single discriminator. This set, T1, produces similar results to set FR1: 2 out of 371 (0.54%) design-set observations are incorrectly assigned. The two sets have five words in common (*in, of, the, all, too* and *now*). In addition to these, set T1 includes *which* and *dare*, for which set FR1 substitutes the more common words *are* and *must*.

As noted in Section 5.5.1, pooling the counts of a set of infrequent markers might produce a useful variable for recognizing either writer. Such a set was chosen for each author, resulting in two variables "Infreq-Fl+" and "Infreq-Sh+." The backward elimination procedure was repeated after including these two variables. Two new sets were produced:

the of Infreq-Fl+ *in* Infreq-Sh+ *all no now* (Set FR2)

(corresponding to set FR1) and:

Infreq-Sh+ Infreq-Fl+ *in of too the dare* (Set T2)

(corresponding to set T1). These sets performed slightly better on the design-set data; set T2 misclassified only one observation and set FR2 correctly assigned all 371.

6.4.3 Dimensionality and Accurate Estimation

In his recent book on density estimation Silverman discusses why the number of samples used in the design set becomes more important when a pdf is estimated in a high-dimensional space [142, pp. 91–94]. The problem of underflow in program KERCON was mentioned earlier. Silverman provides some insights

into why this can be a problem. The characteristics of the univariate normal distribution are very familiar to most students of statistics. But in 10 dimensions some of these characteristics are very different for this well-behaved distribution. For example, regions of very low density become quite important; over half the observations will lie in regions where the pdf is less than one one-hundredth of its maximum. Likewise, large regions of high density may contain very few observations for a sample of moderate size. For the 10 dimensional normal, 99% of the mass of the distribution lies more than 1.6 standard deviations from the origin.

Thus the tails of a distribution become more important in high-dimensional spaces. This is one reason for the recent interest in variable kernel methods, which more accurately estimate the pdf in distribution tails. Of course another implication is that more observations are needed in order to obtain an accurate pdf estimate at any given point in the measurement space. Silverman shows that the number of design-set observations needed for accurate estimation rises quite rapidly with dimensionality. Table 6–4 gives the number needed to estimate the pdf at the origin of a unit multivariate normal distribution in order to achieve a relative mean square error of less than 0.1.

According to the table, to achieve this accuracy in a pdf estimate based on the design-set scenes of either author, only three or four variables could be used. (Only 106 scenes from the 6 plays of the Fletcher design set contain 500 words or more.) However, in the feature selection process described earlier, for forward selection the misclassification rates decreased until 8 variables had been selected. Again, the backward elimination procedures confirmed that using fewer variables increased the number of scenes assigned to the wrong author. While this might seem puzzling in view of Table 6–4, the likely explanation is that the decrease in accuracy is not significant enough to affect the ratio of the Fletcher and Shakespeare estimates. But Silverman's findings cause some unease, especially when the likelihood ratio for a sample is borderline. Partly for this reason, in

For estimating a standard normal distribution at the origin using a normal kernel, this table lists number of samples required to ensure that the relative mean square error is less than 0.1. (From Silverman [142], page 94.)

Dimensions	Num. of Samples
1	4
2	19
3	67
4	223
5	768
6	2790
7	10700
8	43700
9	187000
10	842000

Table 6–4: Number of samples needed for accurate estimation in n dimensions

Section 6.5.4 a reject option will be introduced to recognize scenes that cannot be safely assigned by a classifier to either author.

6.5 Performance on Samples of Known Authorship

As noted in Section 6.3 a classifier can be judged using the leaving-one-out method to allocate the observations from which it was designed. In this study there were good reasons for withholding a number of plays. This test set can be used to confirm the performance measures of the variable subsets selected in the preceding section. The results for the kernel method (which was used in feature selection) are presented first. For each of the variable subsets selected, Table 6–5 summarizes program KERCON's misclassification results for the design and test-set scenes.

As expected the misclassification rate for the test-set observations is larger than that found with the leaving-one-out method. Subsets T1 and T2 perform

slightly better than the other four feature subsets, set T2 failing to correctly identify 4 scenes (4.6%). Set FR2 shows the largest change in the misclassification rate. Although it successfully assigns all 371 scenes in the design set, for the test set it is the least successful of all six subsets, misallocating 11 of 88 (12.5%) samples. As noted earlier a classifier developed using distribution-free methods may reflect the peculiarities of the design-set observations yet fail to "generalize" well enough to accurately classify independent observations. The poorer performance for the two word subsets chosen using forward sequential selection, FS1 and FS2, is apparent. Although their number of misclassifications for the test-set scenes is comparable to sets FR1 and FR2, sets FS1 and FS2 were judged less reliable than the four sets selected using backward elimination (T1, T2, FR1 and FR2). Therefore these two sets were not considered further in the study.

	FS1	FS2	T1	T2	FR1	FR2
Design Set: 222 scenes $\geq$ 1000 words	0 0.0%	1 0.5%	—	—	—	—
Design Set: 371 scenes $\geq$ 500 words	8 2.2%	9 2.4%	2 0.5%	1 0.3%	1 0.3%	0 0.0%
Test Set: 88 scenes $\geq$ 500 words	10 11.4%	8 9.1%	5 5.7%	4 4.6%	8 9.1%	11 12.5%

Note: At this stage of the analysis, only sets FS1 and FS2 were tested in scenes of at least 1000 words. Later in the study, the other 4 sets were tested with samples of various lengths. These results are presented in Table 6–9 on page 277.

Table 6–5: Kernel method misclassifications for feature subsets

Linear Discriminant Function:

	T1	T2	FR1	FR2
Design Set: 371 scenes $\geq$ 500 words	25 6.7%	21 5.6%	29 7.8%	24 6.5%
Test Set: 88 scenes $\geq$ 500 words	7 8.0%	4 4.5%	5 5.7%	7 8.0%

Quadratic Discriminant Function:

	T1	T2	FR1	FR2
Design Set: 371 scenes $\geq$ 500 words	27 7.3%	22 5.9%	31 8.4%	28 7.5%
Test Set: 88 scenes $\geq$ 500 words	8 9.1%	4 4.5%	7 8.0%	10 11.4%

Table 6–6: Misclassifications using the linear and quadratic discriminant functions

Table 6–6 shows that this distribution-free method does out-perform the parametric classifiers that assume normal distributions for these sets of variables. SAS's procedure DISCRIM was again used to classify the design and test sets using the linear and quadratic discriminant functions. For all four subsets of words, the number of misclassifications for the design set is appreciably larger for both functions. The test-set results are not nearly so poor. In fact, for sets FR1 and FR2 the parametric procedures misclassify fewer scenes than the kernel method. This may simply reflect peculiarities of the six plays in the test set.

6.5.1 The Effectiveness of k-NN Classification

The four feature subsets selected using backwards elimination were evaluated using nearest neighbor methods with k ranging from 1 to 9. Table 6–7 compares the design and test-set misclassification rates for these classifiers with the kernel method results. The k-NN classifiers are much less accurate in classifying the design set. The number of incorrectly assigned scenes ranges from 19 (5.1%) up to 54 (14.6%). Again, the choice of value for k makes a great deal of difference, and the best value for one subset of words may not be optimum for another.

The best results for the test-set scenes are only slightly poorer than the kernel results. One puzzling aspect of the k-NN results is the fact that the classifiers quite often show a lower misclassification rate for the test set than for the design set (except for FR2, where once again the performance is noticeably poorer for the test set). As noted in Section 6.3 one expects the holdout method (from classifying an independent test set) to produce a pessimistic estimate of the error rate; the leaving-one-out method should result in a smaller (and more accurate) value.

One striking feature of Table 6–7 is the superior performance for all subsets of words when $k = 7$. This result is due to the nature of the leaving-one-out method and the k-NN classification rule (6.11). By coincidence the numbers of Fletcher and Shakespeare scenes in the design set are in a ratio of 2 to 5. This gives rise to the possibility of a tie if an independent observation has 2 Shakespeare and

For each value of k and each feature subset, values are given showing the number of misclassifications for the 371 design set scenes and the 88 test set scenes. For $k = 7$ the number of ties is indicated in brackets. Corresponding values for the weighted NN and kernel methods are provided for comparison.

Nearest Neighbor classification:

	Feature Subsets							
	T1		T2		FR1		FR2	
k	Design	Test	Design	Test	Design	Test	Design	Test
1	40 10.8%	14 15.9%	45 12.1%	13 14.8%	41 11.1%	11 12.5%	37 10.0%	12 13.6%
2	43 11.6%	8 9.1%	42 11.3%	9 10.2%	46 12.4%	8 9.1%	37 10.0%	12 13.6%
3	48 12.9%	12 13.6%	51 13.7%	11 12.5%	54 14.6%	12 13.6%	45 12.1%	12 13.6%
4	41 11.1%	8 9.1%	29 7.8%	9 10.2%	50 13.5%	8 9.1%	24 6.5%	8 9.1%
5	45 12.1%	8 9.1%	37 10.0%	8 9.1%	45 12.1%	10 11.4%	25 6.7%	9 10.2%
6	42 11.3%	7 8.0%	43 11.6%	8 9.1%	46 12.4%	10 11.4%	30 8.1%	10 11.4%
7	35 9.4%	4(7) 4.5%	27 7.3%	4(6) 4.5%	36 9.7%	6(7) 6.8%	19 5.1%	8(4) 9.1%
8	41 11.1%	7 8.0%	35 9.4%	7 8.0%	47 12.7%	11 12.5%	27 7.3%	9 10.2%
9	42 11.3%	7 8.0%	42 11.3%	7 8.0%	45 12.1%	8 9.1%	30 8.1%	9 10.2%

Weighted Nearest Neighbor classification:

1	82 22.1%	17 19.3%	81 21.8%	19 21.6%	108 29.1%	20 22.7%	99 26.7%	23 26.1%

Kernel estimator:

	2 0.5%	5 5.7%	1 0.3%	4 4.5%	1 0.3%	8 9.1%	0 0.0%	11 12.5%

Table 6–7: Number of misclassifications in the design and test sets using k-nn methods

5 Fletcher neighbors:

$$k_{\mathrm{Sh}}/n_{\mathrm{Sh}} = 5/265 \;=\; k_{\mathrm{Fl}}/n_{\mathrm{Fl}} = 2/106$$

(In Table 6–7, the counts in parentheses for $k = 7$ indicate the number of test-set samples which cannot be assigned because of such a tie.) For the design set, when reclassifying a sample belonging to ω_i, the value of n_i is decremented by 1, thereby upsetting the 2 to 5 ratio. If such a scene has 2 Shakespeare and 5 Fletcher neighbors, the decision tilts towards ω_i, the true class of the sample. Thus, when a Shakespeare scene is left out:

$$k_{\mathrm{Sh}}/n_{\mathrm{Sh}} = 5/(265-1) \;>\; k_{\mathrm{Fl}}/n_{\mathrm{Fl}} = 2/106$$

This peculiar behavior is part of the larger problem of choosing a proper value for k. But even with this advantage, the results for $k = 7$ are not close to the kernel results for either the design or test data.

One major difference between the k-NN methods and the kernel estimation method as implemented in program KERCON concerns the within-class variance of the individual variables. k-NN classification is based on Euclidean distances, and no allowance is made for a greater degree of variation in different dimensions. In contrast, the classifier described by (6.9) on page 245 uses the standard deviation and the smoothing parameter to scale each contribution $y_j - x_{ij}$ for dimension j and each design-set point $\mathbf{x}_i$. The normal kernel function thus has a different shape in each dimension.

The two methods are using the information in the measurement space in a very different manner. One might first imagine that initially standardizing the data to have unit variance might make the two methods comparable. But when estimating the pdf $\hat{p}(\mathbf{x} \mid \omega_i)$, the kernel method allows for different variances within each class. If one decides to initially standardize the measurement space, one will have to use the pooled variances calculated from observations of all classes. (One could transform each class individually, but what values would be used to scale unclassified observations?)

For each value of k and each feature subset, values are given showing the number of misclassifications for the 371 design set scenes and the 88 test set scenes. For $k = 7$ the number of ties is indicated in brackets. Corresponding values for the weighted NN and kernel methods are provided for comparison.

Nearest Neighbor classification:

	Feature Subsets							
	T1		T2		FR1		FR2	
k	Design	Test	Design	Test	Design	Test	Design	Test
1	41 11.1%	11 12.5%	28 7.5%	10 11.4%	44 11.9%	9 10.2%	31 8.4%	15 17.0%
2	36 9.7%	9 10.2%	31 8.4%	9 10.2%	36 9.7%	8 9.1%	32 8.6%	13 14.8%
3	45 12.1%	11 12.5%	39 10.5%	9 10.2%	42 11.3%	11 12.5%	39 10.5%	13 14.8%
4	34 9.2%	8 9.1%	29 7.8%	5 5.7%	37 10.0%	8 9.1%	27 7.3%	10 11.4%
5	32 8.6%	9 10.2%	30 8.1%	5 5.7%	36 9.7%	7 8.0%	30 8.1%	12 13.6%
6	35 9.4%	9 10.2%	32 8.6%	5 5.7%	34 9.2%	8 9.1%	30 8.1%	11 12.5%
7	23 6.2%	7(4) 8.0%	24 6.5%	5(3) 5.7%	29 7.8%	4(4) 4.5%	24 6.5%	10(3) 11.4%
8	35 9.4%	7 8.0%	33 8.9%	6 6.8%	33 8.9%	5 5.7%	27 7.3%	10 11.4%
9	34 9.2%	7 8.0%	35 9.4%	6 6.8%	31 8.4%	7 8.0%	28 7.5%	11 12.5%

Weighted Nearest Neighbor classification:

1	82 22.1%	19 21.6%	80 21.6%	17 19.3%	101 27.2%	20 22.7%	108 29.1%	20 22.7%

Kernel estimator:

	2 0.5%	5 5.7%	1 0.3%	4 4.5%	1 0.3%	8 9.1%	0 0.0%	11 12.5%

Table 6–8: Misclassifications using k-NN methods with standardized data

Dividing the observations for each variable by the pooled standard deviations will have no effect on results using KERCON's kernel classifier. (In fact, from (6.9) one can see that multiplying all the values of one or more variables by a constant will not alter the pdf estimates.) On the other hand, standardizing the measurement space generally improves the k-NN results somewhat. These are listed in Table 6–8. The improvement for set T2 and $k = 1$ is striking, although the NN results for the other sets are hardly affected. Again, the classifiers' performance on the test-set samples is often better than on the design set, especially for sets T2 and FR1. But the poor test-set performance of set FR2 appears to be exacerbated by standardization.

In any statistical study the analyst must decide whether or not to transform the data to have unit variance. Statistical textbooks usually state that it should be considered if the data values are different in nature or scale. This is not really the case for these word-rate variables. Dividing each observation by the standard deviation calculated by pooling data from both authors seems difficult to justify; inspection of the statistics listed in Table 5–12 on page 215 suggests that the authors' standard deviations are different for many words. Initial transformation seems a crude option. If one decides to recognize the different variance of individual variables, explicitly incorporating this factor into pdf estimation is sensible.

Even using a standardized measurement space, the k-NN method's performance is clearly inferior to the kernel method for this data. The kernel method's ability to allow for different within-class variances for each variable appears to be an important factor. The k-NN results (particularly the leaving-one-out method) may also be complicated by the different number of observations in the design-set classes, although the development leading up to (6.11) appears to include this factor. (Using a reduced number of Shakespeare samples in the design set and word set T2, the kernel method still out-performs the k-NN classifiers when the number of samples in the design-set classes are equal.) A final problem that has not resolved itself is an objective procedure for determining the best

choice of value for k. For these reasons, the nearest neighbor methods will not be used in further analyses. This decision was taken with some regret, since the basic principle is relatively simple and intuitively appealing. But it seems senseless to apply the methods to disputed data without being able to evaluate their effectiveness on classified samples.

6.5.2 Characterization Effects

The plays in the test set can also be used to determine whether these writers' characterization skills can affect classification based on these marker words. A computer program was used to identify and group speeches according to speaker. (Speech assignments were primarily based on the speech headings in machine-readable versions of the 17th century texts. However, the critical apparatus of the Bevington edition of Shakespeare and the Bowers editions of Fletcher were examined, and the assignments modified according to these editors' emendations.) The samples corresponding to characters speaking at least 500 words were then used as a test set with each of the four feature subsets. The following table shows the misclassification rates for these 62 samples, comparing them with the results for complete scenes.

	T1	T2	FR1	FR2
Test Set: 88 scenes $\geq$ 500 words	5 5.7%	4 4.6%	8 9.1%	11 12.5%
Test Set: 62 speakers of at least 500 words	4 6.5%	3 4.8%	6 9.7%	6 9.7%

When compared to those for the test set divided into scenes, the values are only slightly higher overall for three sets and somewhat better for set FR2.

Further examination of function word rates and characterization in Jacobean dramas might produce a great deal of valuable information. In the current application, it needs to be established that the traits being measured in a sample of text reflect authorship and are not associated with the characters present in the

sample. Burrows has studied modal auxiliaries, pronouns and other common word-classes in the novels of Jane Austen [22]. Using sophisticated statistical methods he demonstrates these can be used to differentiate and group characters in Austen's works. It need not be disturbing that his research discovers significant differences between characters using variables that are very similar to those being used in this study. As long as the internal variation (no matter what its source) is smaller than the differences between Shakespeare and Fletcher, discrimination is possible. The fact that misclassification did not increase a great deal when the test-set plays were re-divided according to speaker indicates that, for the purpose of recognizing authorship, these marker words are relatively immune to characterization effects in these plays.

6.5.3 Sample Length and the Misclassification Rate

The four sets of marker words under examination have been selected and tested using design and test sets made up of scenes of at least 500 words. At this point it is informative to examine their performance when different minimum-length criteria are used. Table 6–9 lists the number and percentage of misclassifications when the kernel method is used with acts and with scenes containing at least 1000, 750, 500, 400 and 300 words.

There is probably a trade-off involved in any decision to include shorter scenes in the design set. Using more samples should lead to a better classifier. But as noted in the last chapter, the variance of word-rate variables increases when shorter scenes are examined. While the design-set results do not deteriorate all that much, the increasing misclassification rates for the test set may indicate that the latter effect dominates the former. The 500 word minimum appears to be a good choice for examining the disputed scenes of *Henry VIII* and *The Two Noble Kinsmen.* Of the short scenes in these two plays, only two contain between 400 and 500 words (one of them 497); two more contain between 300 and 400 words. Therefore little could be gained at the risk of introducing more uncertainty into

Design Set:	T1	T2	FR1	FR2
131 acts	0 0.0%	0 0.0%	0 0.0%	0 0.0%
222 scenes $\geq$ 1000 words	0 0.0%	0 0.0%	0 0.0%	0 0.0%
291 scenes $\geq$ 750 words	2 0.7%	1 0.3%	1 0.3%	0 0.0%
371 scenes $\geq$ 500 words	2 0.5%	1 0.3%	1 0.3%	0 0.0%
411 scenes $\geq$ 400 words	3 0.7%	4 1.0%	4 1.0%	1 0.2%
458 scenes $\geq$ 300 words	4 0.9%	5 1.1%	2 0.4%	2 0.4%

Test Set:	T1	T2	FR1	FR2
30 acts	1 3.3%	1 3.3%	4 13.3%	1 3.3%
50 scenes $\geq$ 1000 words	1 2.0%	1 2.0%	1 2.0%	7 14.0%
66 scenes $\geq$ 750 words	3 4.5%	1 1.5%	5 7.6%	8 12.1%
88 scenes $\geq$ 500 words	5 5.7%	4 4.5%	8 9.1%	11 12.5%
100 scenes $\geq$ 400 words	8 8.0%	6 6.0%	11 11.0%	14 14.0%
113 scenes $\geq$ 300 words	12 10.6%	9 8.0%	17 15.0%	19 16.8%

Table 6–9: Misclassification rates using samples of different length

the classifiers. These 371 design-set scenes account for 92.6% of the total number of words in the 6 Fletcher plays and 92.1% of those in the 20 Shakespeare plays.

Set FR2 performs much more poorly on the independent test-set observations even when scenes of less than 1000 words are excluded. This set mainly differs from the other three sets by including *no* and excluding *too*. These two differences appear to be critical, and at this point it seems in order to exclude set FR2 from further consideration. While the performance of set FR1 is not as good as sets T1 and T2, this set is retained. All three sets of words contain many of the same words, but FR1 is different enough from the other two (for example, it contains *are* and *must*, but excludes the infrequent marker *dare*) that it may prove useful as a (somewhat) independent check.

6.5.4 Implementing a Reject Option

Earlier it was noted that Silverman's results bring into question the accuracy of kernel estimation when based on only 106 Fletcher and 265 Shakespeare scenes. The high proportion of correctly classified scenes in the design and test sets indicate that this may not often seriously affect assignment. But this problem encourages the use of a reject option (described in Section 6.1.3, page 238). By only accepting for classification samples for which the likelihood ratio is larger than some value, one hopes to recognize observations near the decision surface, where inaccurate density estimates might have the most serious effect. Clearly some scenes will be rejected that might otherwise be classified correctly. A reject threshold is an objective way of implementing a policy of caution that asserts that incorrect classifications are less desirable than letting some observations remain unclassified.

The three subsets of words (T1, T2 and FR1) were used to classify the design and test sets (scenes of 500 words or more) using different threshold values t. (Recall that an observation is accepted for classification only if the posterior probability is greater than $1 - t$.) Table 6–10 lists the number of scenes that were misclassified and the number rejected. Note that a very large majority of the

For each value of t and each word set, the first count is the number of misclassified scenes; the second is the number of scenes rejected. (The proportion of the total number of observations in each set is given below each count.)

Design Set (371 observations):

t	T1		T2		FR1	
0.40	2 0.5%	2 0.5%	1 0.3%	3 0.8%	1 0.3%	5 1.3%
0.35	0 0.0%	5 1.3%	1 0.3%	5 1.3%	1 0.3%	7 1.9%
0.30	0 0.0%	8 2.2%	0 0.0%	10 2.7%	1 0.3%	14 3.8%
0.25	0 0.0%	12 3.2%	0 0.0%	13 3.5%	0 0.0%	20 5.4%
0.20	0 0.0%	14 3.8%	0 0.0%	18 4.9%	0 0.0%	26 7.0%

Test Set (88 observations):

t	T1		T2		FR1	
0.40	5 5.7%	2 2.3%	4 4.5%	1 1.1%	5 5.7%	3 3.4%
0.35	5 5.7%	4 4.5%	4 4.5%	3 3.4%	5 5.7%	4 4.5%
0.30	4 4.5%	6 6.8%	3 3.4%	4 4.5%	4 4.5%	6 6.8%
0.25	3 3.4%	8 9.1%	3 3.4%	7 8.0%	4 4.5%	9 10.2%
0.20	3 3.4%	8 9.1%	3 3.4%	9 10.2%	4 4.5%	12 13.6%

Characters in the Test Set (62 observations):

t	T1		T2		FR1	
0.20	2 3.2%	7 11.3%	0 0.0%	8 12.9%	4 6.4%	10 16.1%

Table 6–10: Using a reject option: the number of misclassified and rejected samples

design-set scenes are classified correctly with large posterior probabilities. Even for the worst performer, set FR1, 345 of 371 scenes (93.0%) have probabilities greater than 0.80. Those scenes in the design set that were initially misclassified are rejected when $t = 0.25$.

The test-set results are not as good. Several scenes are misclassified with probabilities larger than 0.80 for at least one of the word sets. The proportion of test-set scenes that are correctly classified (that is, not rejected or misclassified) ranges from 87.5% for set T1 to 81.8% for set T2. Samples composed of the speeches of characters in the test-set plays were also tested with $t = 0.20$. The percentages of misclassified and rejected samples are slightly higher compared to the corresponding values for scenes from the test set. These results support the conclusion of Section 6.5.2; characterization in the test set does not change the rate of occurrence of these markers enough to cause much concern about the procedure.

If a scene is assigned to Shakespeare and $t = 0.25$, then the Shakespeare pdf estimate is at least 3 times the Fletcher pdf estimate. A value $t = 0.20$ corresponds to a ratio of 4 to 1. A threshold of 0.20 is probably unusually strict for an application of discriminant analysis. However, caution in an attribution study based on word rates is probably not unwise. Therefore I decided to use this value ($t = 0.20$) in the classification of disputed scenes. Before applying the classifiers to scenes from *Henry VIII* and *The Two Noble Kinsmen* the misclassified and rejected scenes in the design and test sets must be examined. Also, a procedure for making a decision where the different subsets of markers produce different results must be formulated.

6.5.5 Examination of Misclassified and Rejected Scenes

One can postulate two explanations for the failure to correctly classify a scene. Any classification procedure is only as good as the variables it uses. A glance at Figure 6–1, the graph of the rates of *in* and *of* in scenes, shows that the observations for each author do not lie in well-separated clusters. While discrimination

is improved by using more variables, it should not come as a surprise that some observations are misclassified, even when a large number are examined. The second explanation stems from the nature of words and language. A sample of text may include an unusually large or small number of occurrences of one or more marker words, perhaps with no obvious stylistic explanation. As noted in Section 3.1.4, in some samples one can recognize that a local stylistic effect is responsible for the spurious result. This is especially true for short samples or when low-frequency features are involved.

Although the kernel estimation method uses the distances to the classified design-set observations to calculate posterior probabilities of authorship for a unclassified sample, one would expect that differences between the sample's word rates and the mean rates of the two authors might be reflected in the result. To show this, rates for three "new" observations were calculated. The first observation's word rates were set equal to the mean rate of the Shakespeare design-set samples for all the markers, and the second observation's rates were assigned the Fletcher mean rates. Each word rate for the third observation was set to the mid-point between the two author's average rate. Program KERCON classified the 3 samples for all three subsets of words:

	Posterior Probabilities		
All word rates =	T1	T2	T3
$\overline{x}_{\mathrm{Fl}}$	Fl 0.958	Fl 0.962	Fl 0.976
$\overline{x}_{\mathrm{Sh}}$	Sh 0.995	Sh 0.995	Sh 0.985
$(\overline{x}_{\mathrm{Fl}} + \overline{x}_{\mathrm{Sh}})/2$	Fl 0.554	Fl 0.709	Sh 0.630

Happily, the procedure classifies the first two observations as one would expect. For both observations, the posterior probabilities for all three subsets are greater than 95%. The third observation would be rejected for all subsets using a threshold $t = 0.20$. The probability for Set T2 confirms the advisability of using a reject option. If only the best subset of words (T2) had been retained and a more tolerant reject threshold employed, an observation with these word rates would be assigned to Fletcher.

The difference between a sample's rates and the two authors' mean rates can be used to gain some insight into why some scenes are not classified correctly. A program was written that prints (for each author) each variable's value and its deviation from the author's mean $\overline{x}$ in terms of the within-author standard deviation s. This statistic is often referred to as a *z-score:*

$$z = (x - \overline{x})/s \tag{6.12}$$

If one word has an abnormally high rate in a sample, then this sometimes results in a large z-score for both authors. (If the rates were distributed normally around the mean, the 95% significance level is 1.96.) This appears to be the case in some of the misclassified or rejected scenes, usually involving one word that the true author normally uses less often than the other. In many cases the word involved is a more frequent marker (for example, *in, of* or *the*). In such situations it is difficult to determine what (if any) stylistic peculiarities were responsible for the unusually high rate. If the word is less frequent, an explanation for the anomalous usage can sometimes be proposed.

The results for any scene that is misclassified or rejected (for $t = 0.20$) by any of the sets T1, T2 and FR1 were examined in detail. In this examination z-scores often proved useful, but sometimes no conclusion could be drawn from their values. Again, z-scores can give an indication that a particular word is greatly affecting a result, but they obviously cannot "explain" the kernel classifier's result. They are not multivariate and do not consider the classification of other scenes with similar rates.

For the design set, 41 of 371 scenes (11.1%) are rejected and none misclassified. Of these 41, a large majority (28) are rejected by only one word set. Five (1.4% of the total number of scenes) are rejected by all three word sets and eight (2.2%) by two sets. First, the five scenes in the design set are rejected by all three word sets will be discussed.

Bond I.i (1591 words) has high rates for two words favored by Shakespeare, *of* ($z_{\mathrm{F1}} = 1.8$) and *the* ($z_{\mathrm{F1}} = 1.8$), for no obvious reason. No individual word or

words seem responsible for the rejection of *Deme* II.iii (1047 words). One fairly short scene, *Priz* I.iv (647 words), has low rates for most of the markers; this scene is marked by short exchanges and questions. Another scene from *Priz*, III.i (869 words), has a large number of occurrences of *which* ($z_{\mathrm{Fl}} = 4.6$) but is also rejected by sets T2 and FR1 which do not include this word. No other individual word has a striking z-score. A repetition of four occurrences of *too* in four consecutive lines of *TGV* I.ii (lines 92–95) produces a Shakespeare z-score of 2.7:

> *Lu.* Keepe tune there still; so you will sing it out:
> And yet me thinkes I do not like this tune.
> *Iu.* You do not?
> *Lu.* No (Madam) it is too sharpe.
> *Iu.* You (Minion) are too saucie.
> *Lu.* Nay, now you are too flat;
> And marre the concord, with too harsh a descant...

Perhaps one could consider this example a positive indication of the method's robustness to repetitions of the Fletcher marker *too*. Since the three word sets yield posterior probabilities between 0.66 and 0.76 for Shakespeare, the scene could not have been assigned to Fletcher even with a less severe reject threshold.

Eight other design-set scenes are rejected by two of the word sets. For all but one of these eight, the three sets would have agreed on the correct author if no reject option had been employed. This "low-but-agreeing" situation also holds for all 28 of the scenes that are only rejected by one set of words. All three sets agree on the correct author, but the posterior probability for one set is less than 0.80. This suggests that one might wish to accept such scenes for classification despite the fact that one subset of words produces a likelihood ratio of less than 4 to 1.

Such a policy, however, results in two misclassifications for the 13 scenes in the test set where only a single result does not meet the acceptance criteria. For 2 of these 13, *Tem* III.ii (1161 words) and *Vale* I.iii (2093 words), the two acceptable probabilities do not agree on which author wrote the scene; thus these

cannot be assigned. Results for the other eleven scenes would agree on a single author if no reject option were in effect, but the results for two of these point to the wrong man. The classification of Fletcher's *Thom* V.i (620 words) appears to be affected by a high rate for *in* (11 observed occurrences, 5.2 expected, $z_{\mathrm{Fl}} = 2.8$). Also, there are only 6 occurrences of *the* in the scene, fewer than would be expected for either author ($z_{\mathrm{Fl}} = -1.9$, $z_{\mathrm{Sh}} = -2.3$). No quirks of style are evident that might explain these rates. The posterior probabilities for sets T2 and FR1 are strong results for Shakespeare (0.93 and 0.99, respectively), while the rejected T1 result is 0.66.

The other test-set scene that would be misclassified if one rejection were allowed is *Ant* IV.xv. Of 719 total words, this scene contains 3 occurrences of *dare*, including Cleopatra's repetition in consecutive lines of "I dare not." The word *dare* is used more frequently by Fletcher than Shakespeare; the latter's mean rate of occurrence is only 0.26. These 3 occurrences result in a rate of 4.2 and a Shakespeare z-score of 7.0. Here a small number of additional occurrences has a tremendous effect on the value of the variable. This demonstrates the desirability of measuring frequent features, especially when converting word counts to word rates.

In this scene this use of *dare* is combined with a high rate of *now*, a more frequent word also favored by Fletcher. The scene contains 7 occurrences, several found in Antony's dying speech. Antony's death concentrates attention on the present situation, and the increased use of *now* results in a rate of 5.0 and a Shakespeare z-score of 3.2. While one can propose a stylistic explanation for the increased use of these uncommon words in Shakespeare, an effective method for studying authorship should be unaffected by conscious stylistic manipulation. The number of correct classifications show that this is *usually* the case, but scenes such as *Ant* IV.xv remind one that this goal may never be completely achieved.

These two scenes would be the only ones misclassified in either the design or test set (and only if one low posterior probability were accepted). Most other

scenes that are incorrectly classified by one of the sets of words are rejected by another set. In addition, for three scenes the posterior probability for each word set is greater than 0.80, but the classification results conflict. Fletcher's *Vale* V.viii (965 words) has large numbers of *of* ($z_{\mathrm{Fl}} = 2.7$) and *the* ($z_{\mathrm{Fl}} = 1.5$), but the pooled sets of infrequent markers suggest the correct author. Thus T2 assigns this scene to Fletcher ($p = 0.805$) but T1 and FR1 incorrectly classify it as Shakespeare ($p = 0.93$ and $p = 0.98$). Act IV Scene xiv of *Antony and Cleopatra* (1144 words) is incorrectly assigned to Fletcher by set FR1 with $p = 0.87$, apparently because of the 10 occurrences of *now* ($z_{\mathrm{Sh}} = 2.7$). (Note that this scene precedes the misclassified *Ant* IV.xv, which contains too many occurrences of *now* and *dare*.) Finally, *Tem* III.iii (922 words) has a remarkably high rate for *are*; 16 occurrences produce a rate of 17.35. Such a rate is untypical for both writers but more so for Shakespeare ($z_{\mathrm{Fl}} = 3.1$ and $z_{\mathrm{Sh}} = 4.8$) since his average is lower. Examination of the text reveals no repetitions or patterns that might help explain this.

Two scenes in the test set are rejected by all three sets of words. *Richard III* II.iv (598 words) has low rates for *in* and *of*. The other is yet another scene from the third act of *The Tempest:* Scene i (827 words), which may be affected by high rates for *are* and *dare* (only one occurrence produces a Shakespeare z-score of 1.7) and a low rate for *the* ($z_{\mathrm{Sh}} = -1.5$). Finally, four scenes are rejected by two of the three word sets. One of these provides an excellent illustration of how a stylistic effect can influence classification.

In *As You Like It,* V.ii and the collected speeches of Silvius are misclassified or rejected for sets T1 and FR2. This is due to an extremely large number of occurrences of *all* in the shepherd's description of "what 'tis to love:"

It is to be all made of sighes and teares,
And so am I for Phebe...
It is to be all made of faith and seruice,
And so am I for Phebe...
It is to be all made of fantasie,
All made of passion, and all made of wishes,

> All adoration, duty, and obseruance,
> All humblenesse, all patience, and impatience,
> All puritie, all triall, all obseruance:
> And so am I for Phebe.

The Shakespeare z-score for this scene is 3.7. Set T2, which does not include *all*, correctly classifies the scene with a posterior probability of 0.95. This is a striking example of how a function word can become part of the subject matter and lose its usefulness as a subconscious marker of authorship. This example also supports the retention of all three sets of marker words for use in evaluating disputed samples rather than choosing the single best performer. The slight differences between sets may produce ambiguous results that can be traced to such an unusual usage.

In summary, when a reject threshold of 0.20 is used, 330 of 371 (88.9%) design-set scenes of at least 500 words are accepted for classification by all three sets T1, T2 and FR1. For the test set, 64 of 88 scenes (72.7%) are accepted and assigned to the same author by all three sets of words. None are misclassified in either the design or test sets, a combined accuracy rate of 88.7%. Three scenes of the test set (3.41%) have posterior probabilities greater 0.80 for all three word sets, but the results do not agree. For 13 design-set and 6 test-set scenes, the probabilities for at least two of the sets of words are less than 0.80.

If one is willing to classify a sample if a single set's posterior probability is less than 0.80, then 28 more design-set and 13 more test-set samples are classified. (For 2 other scenes in the test set, the assignments for the two un-rejected probabilities disagree.) This results in a 96.5% acceptance rate for the design set, again with no misclassifications. For the test set, 87.5% of the scenes are now accepted for classification, but two of these are misclassified. The resulting error rate is 2.6% of those accepted for classification or 2.3% of the total number of test-set scenes. The confidence interval associated with this value is $(0\%, 5.4\%)$, according to the formula described in Section 6.4.2. Again combining the results for the design and test sets, 435 of the 459 scenes are accepted and classified, with only the two misclassifications. This yields a combined classification rate

of 94.77%, a misclassification rate of 0.44% and a rejection rate of 4.79%. (The procedure described in this paragraph was adopted for the assignment of scenes in *TNK* and *H8*.)

6.5.6 Scenes of Joint Composition

It is possible that some individual scenes in *Henry VIII* and *The Two Noble Kinsmen* are not by a single writer, but represent the joint work of both Fletcher and Shakespeare. More detailed discussion of this possibility will be postponed until the next chapter, but it might prove valuable to see how the procedures developed in this chapter assign samples known to be of joint composition. These samples will be "manufactured" from scenes in the test set. While the primary reason for introducing a reject option stems from an awareness of possible inaccuracies in pdf estimation, it was hoped that scenes of joint authorship might often be rejected or ambiguously assigned by the three word sets.

"Joint composition" includes numerous possibilities, of course, ranging from one author just touching up another's work in a few lines to a complete overhaul of an existing scene. This examination will be limited to scenes composed of roughly the same number of words by Shakespeare and Fletcher. This might model a situation where different dramatists were responsible for different exchanges or episodes within a single scene. It was also considered desirable to study a number of samples of approximately the same length. Twenty pairs of scenes (of at least 400 words) from the test-set samples of the two authors were chosen using a random number table. The first 400 words or so (about 65 lines of verse) of each scene were used to create a joint sample of between 800 and 900 words (an intermediate length for a scene). Some variability was introduced because complete speeches were taken when selecting a section of a scene. Word counts were made and program KERCON used with a reject threshold of 0.20 to calculate posterior probabilities of authorship for these 20 "new" collaborations.

Only 9 of the 20 samples (45%) are rejected by one or more of the word sets. Six (30%) are accepted and assigned to the same author by sets T1, T2

and FR1 (2 to Fletcher, 4 to Shakespeare). For five of the samples, each of the three probabilities is larger than 0.80 but two of the classifications disagree on the author. This proportion of high but conflicting probabilities (25%) is much larger than is evident in the test set, where only 3 of 88 scenes (3.4%) had this combination of results. This somewhat supports the hope that ambiguous results may indicate collaboration within a scene.

Of the 9 samples rejected by at least one word set, five are rejected only by a single set. For four of these the assignment with the low probability agrees with the other two, and the sample would be assigned to one of the playwrights if the decision is taken to accept such samples for classification. The fifth of these samples has probabilities of 0.96 and 0.99 for Shakespeare and 0.78 for Fletcher. Finally, three samples are rejected by two of the word sets and another by all three.

In summary, 30% of these scenes are unambiguously assigned to one of the authors. This number is increased by another 20% if samples with only a single rejected probability are accepted for classification. A quarter of the samples result in three posterior probabilities that are greater than 0.80 but for which the assignments do not agree on an author; these and the remaining 25% are rejected. Only limited conclusions can be drawn from these tests. It is certainly not encouraging that one-half of the joint samples can pass as the work of one writer or the other. On the positive side, a comparison with the test-set results indicates that disputed scenes with high but conflicting probabilities might well be made up of the words of both playwrights.

6.6 Summary

Distribution-free discriminant analysis is an attractive multivariate procedure for analyzing word-rate data. The results of this chapter show that these techniques can be used successfully to classify small samples of text taken from plays by Shakespeare and Fletcher. Using some subsets of the sixteen words initially

isolated as good markers of authorship, the kernel method correctly classifies a high percentage of scenes of known authorship. The k-NN classifiers perform more poorly than the kernel methods. This appears to be due to the kernel method's ability to allow for different within-class variation for each variable.

Further analyses of the procedure are encouraging. Division of the test-set scenes according to speaker has little effect on classifier performance. In allowing for possible inaccuracies in the estimation of class-conditional pdfs, a reject option can be used to recognize observations that might be incorrectly classified. Using a strict threshold (but classifying observations for which only one probability does not meet this threshold), 96.5% of the design-set observations are accepted and correctly classified. Over 87% of the test-set samples are accepted, but two of these scenes are incorrectly allocated. The application of the classifiers to samples composed of text taken from both writers' known work indicates that scenes of mixed composition may not always be recognized as such. However, these word-rate variables and statistical techniques appear to be a useful tool for evaluating the authorship question posed by *Henry VIII* and *The Two Noble Kinsmen.*

Chapter 7

Applying the Classifiers to the Disputed Plays

As noted in Chapter 1, there is some external evidence that Shakespeare did indeed collaborate with the younger Fletcher. This evidence centers around *The Two Noble Kinsmen*, but the Stationer's Register, in which London publishers were required to register any book before publication, contains another reference to a joint composition by Shakespeare and Fletcher. In 1653 (37 years after Shakespeare's death) a lost play entitled *The History of Cardenio* was entered by Humphrey Moseley with the pair of men listed as authors. At the same time Moseley also entered the titles of several other plays as Shakespeare's, but these claims of authorship have been unanimously rejected. *Cardenio*'s existence is supported by records of court performances in 1612 or 1613 and by Theobald's publication in 1728 of a play entitled *Double Falsehood*, purportedly revised from Jacobean manuscripts. (Muir discusses *Cardenio* and Theobald's play in *Shakespeare as Collaborator* [114].)

While the external evidence for the two dramatists' partnership in *TNK* has high authority, much of the discussion surrounding this play's authorship concerns the internal evidence it presents. This chapter will examine the authorship question of these two disputed plays in more detail. For both plays, the external evidence pertinent to the authorship question will be reviewed along

with scholars' findings regarding the nature of the sources behind the copy text. Some previous authorship studies based on internal evidence will be discussed,[1] particular attention being paid to quantitative studies of textual features.[2] Finally, rates for the marker words in each scene will be examined using the kernel method, and the results interpreted. Each play will be discussed in turn, beginning with *TNK*, since for this play there is external evidence for collaboration. Some critics have accepted that Shakespeare and Fletcher are jointly responsible for this play and then have claimed that this fact lends support to the possibility of collaboration in *H8*, despite the lack of external evidence.

7.1 *The Two Noble Kinsmen*

An entry for *The Two Noble Kinsmen* was made in the Stationer's Register on 8 April 1634. The publisher was John Waterson, and Fletcher and Shakespeare are listed as joint-authors. Later that year Waterson printed the only quarto of the play, with a title-page noting that it had been: "Presented at Blackfriers by the Kings Maiesties servants, with great applause: Written by the memorable Worthies of their time; Mr. *John Fletcher,* and Mr. *William Shakespeare Gent.*" Title-page ascriptions are not completely reliable, but there are good reasons for accepting this one's accuracy. Waterson appears to have had a good working

[1] The annotated bibliography at the end of Erdman and Fogel's *Evidence for Authorship* [33] thoroughly documents studies of the authorship problem in these two plays through the early 1960s.

[2] The differences between the poetry of Shakespeare and Fletcher will be generally be neglected in this examination. Such differences certainly exist, and a reader unfamiliar with them might wish to consult the introductions to some modern editions of either *TNK* or *H8*. (I recommend Humphrey's New Penguin edition of *H8*. The editor might wish to treat "the vexed question of authorship" as a "tiresome sideline," but he illustrates these differences with verse taken from both writers' undisputed works in addition to *TNK* and *H8*.) Not to recognize these differences in ascribing authorship is to ignore evidence, but some of texts' disputed scenes are made up of prose or verse in which these styles are not clearly evident. And (as always) critics' subjective analyses of poetic style have led to divergent opinions, and an objective analysis of linguistic features is required.

relationship with Shakespeare and Fletcher's company, the King's Men. He published a number of their texts that have undisputed title-page ascriptions.

Indeed, it appears that the source copy behind the 1634 Quarto was obtained directly from the King's Men. The text shows signs of the theater. In three instances, a reminder regarding stage properties is printed in the margin; in each case these anticipate the action of the play by twenty or thirty lines. In addition, two actors' names appear in stage directions. These two men have been identified and only worked with the King's Men during a brief period around 1625–1626. While these and other features suggest a prompt-book, other characteristics point to a manuscript by the author (or authors) or a faithful transcript of such a document. This manuscript would have been annotated by the prompter before a prompt-book was prepared and eventually sold to the publisher (as it was less valuable to the company than the prompt-book itself). The support for this theory is well-documented in Leech's edition of the play [39].

The earliest performance of the play was almost certainly 1613–1614. The morris dancers of III.v are clearly borrowed from Beaumont's *Masque of the Inner Temple and Gray's Inn* which was presented before King James at Whitehall on 20 February 1613. (Some have conjectured that the presence of the characters from Beaumont's masque indicates that he might have had a role in the sub-plot.) Allusions in Jonson's *Bartholomew Fair* indicate that *TNK* would have been familiar to an audience before the end of October 1614, when Jonson's play was first performed. A later court performance in 1619 is suggested by a Revels Office note which mentions the play. The actors' names in the Quarto text indicate a revival in 1625 or 1626.

There is certainly a possibility that the text of the play as we have it reflects changes made for either revival. Proudfoot, in the introduction to his recent edition of the play [38], wonders if Massinger (who succeeded Fletcher as the principal playwright for the King's Men after his death from the plague in 1625) might have written the prologue and epilogue and touched-up the manuscript for the mid-1620s revival. However, he regards this as "far from certain." While

this possibility should be noted, if the theory regarding the nature of the source manuscript is correct, then for the most part the Quarto should reflect the original manuscript. In any case, the 1634 edition is a very good text, with none of the obvious signs of corruption that are evident in *Pericles, Timon* or some other Shakespeare quartos.

In evaluating the external evidence for Shakespeare's participation in *TNK*, one must consider the fact that Heminges and Condell, the publishers of the 1623 Shakespeare Folio, did not include the text in that volume. Hinman's detailed study the Folio [49] indicates that last-minute changes were made with regard to the inclusion of *Timon* and *Troilus and Cressida.* Clearly the selection of some texts for publication of the Folio involved difficulties. So the theory that the text was unavailable to Heminges and Condell at the time is certainly an acceptable possibility. Since these two men were associates of Shakespeare and Fletcher in the King's Men during 1613–14, they were in a very good position to know if either *Henry VIII* or *The Two Noble Kinsmen* were collaborative efforts. However, whether they knew the details of the extent of any collaboration, and how they might have dealt with a play in which Shakespeare had a minor share (he is usually credited with only about a third of *TNK*) can only be conjectured.

7.1.1 Past Studies of Internal Evidence

Two important stylistic analyses of *The Two Noble Kinsmen* were published in the second quarter of the 19th century. Both Spaulding's 1833 article and its review by Hickson in 1847 [48] are founded completely on a subjective analysis of style.[3] Hickson sees evidence of two writers with "dissimilar and unequal powers" and on this evidence divides the play between them. This division was later supported by metrical tests applied by Furnivall and Fleay (and is basically still the accepted division, despite the general discredit heaped upon these last

[3]Hickson's article was reprinted in 1874 in the *Transactions of the New Shakespeare Society*. All references are to this version of the paper.

two disintegrators). The most complete scene by scene analysis of the stylistic and metrical evidence presented in these early studies is provided in Littledale's two-volume edition of the play [37], published in 1876 and 1885.

Fletcher has always been accepted as the major contributor, and the question has centered on whether or not Shakespeare's hand is present. Farnham's study of contracted forms [35] (described in Section 2.2.2) indicates that Massinger almost never used some contractions that occur in the works of Shakespeare and Fletcher (such as *in't, o'the* and *on's*). Another study that further convinced scholars that Massinger was not a major partner in the *TNK* was Hart's study of the play's vocabulary [46]. Shakespeare proves to be a great linguistic innovator, and the parts of *TNK* assigned to him are characteristic of his practice in his later plays. Hart's very detailed examination has met with a great deal of critical approval, although it is certainly open to questions of subjectivity.

To demonstrate the presence of Shakespeare's vocabulary in *TNK*, Hart selects about 1000 of the "rarer" words from the text (about one-third of the vocabulary), keeping note of whether the occurrences fall in the parts attributed to Shakespeare or Fletcher. Comparison of these lists to both a Shakespeare concordance and to the *New English Dictionary* shows that one part conforms to Shakespeare's habit of introducing words new to his own vocabulary and to the English language itself. The vocabulary demonstrated by the author of the other part is "almost entirely derivative."

Hart's identification of a "rare" word is based on his "experience with the vocabularies of nearly 80 plays," and one wonders if words such as *helmeted* and *black-haired* are really significantly rare. However, a few computer searches of each author's word-list generally support Hart's judgement regarding several of the forms he claims are frequently introduced by Shakespeare. For example, neither of the above two words occurs in any of the 33 other plays studied. Shakespeare does indeed appear to use many more "unusual" words ending in the suffixes *-like* and *-less* than Fletcher. One hopes that Hart's findings might be re-examined using more samples of English Renaissance text with the help

of computers and statistical analysis. The variation of such linguistic innovation within individual plays was not examined by Hart. This question is central to an examination of collaboration. Nonetheless, Hart's study provides good evidence of Shakespeare's presence in *TNK*.

An interesting approach to authorship questions revolves around the characteristic use of imagery. Muir has used the idea of *image clusters* in an attempt to recognize Shakespeare in *TNK* [114]. But as noted in *Evidence for Authorship* [33] and by Proudfoot [38, p. xviii] this method has met with some criticism, and Muir's evidence is unconvincing.

Perhaps the most important contribution this century is the examination of *TNK* and *H8* which concludes Hoy's series of articles entitled "The Shares of Fletcher and his Collaborators in the Beaumont and Fletcher Canon" [55]. The first step in his study was the recognition of particular linguistic features that characterizes Fletcher's unaided work (discussed in Section 2.4.1, page 43). Hoy then analyzes their occurrence in a large number of collaborative works. In his analysis of *TNK*, Hoy is quick to point out that the attribution of the rest of the text to Shakespeare must be based on other (that is, non-linguistic) grounds. "Shakespeare uses no language forms which, either in themselves or by virtue of their rate of occurrence, can serve to point immediately and unmistakably to his presence in a play of doubtful authorship." This is rather a sweeping statement, but it does apply to the sort of features with which Hoy is concerned. He does demonstrate that the rate of occurrence of his variables in the non-Fletcherian portions of *TNK* is consistent with Shakespeare's late usage.

Hoy was not the first to recognize the discriminating power of the features he counted, but he was the first to analyze all of them in such a large number of texts. The linguistic evidence is as follows:

1. Shakespeare generally avoids the pronomial form *ye,* for which Fletcher shows a great and unusual preference. All the occurrences in *TNK* fall in scenes that Hoy attributes to Fletcher. (Although there are fewer occurrences of *ye* than might be expected, this can be attributed to alterations

introduced in scribal transmission if a copy of the authors' papers were annotated by the prompter.)

2. The third-person forms of the auxiliaries *hath* and *doth* occur regularly in Shakespeare's works, but Fletcher almost always uses the newer forms *has* and *does*. The one occurrence of *doth* is in the first scene of the play (agreed to be Shakespeare's). Three of the sixteen occurrences of *hath* fall in the Fletcher portions.

3. Shakespeare's use of *'em* never exceeds Fletcher's although it is almost equal in *The Tempest* and *Timon*. Several other contracted forms (*i'th'* and *a'th'/o'th'*) are also used at different rates but are not as distinctive. The use of these in the division tested by Hoy also conforms to the authors' observed pattern of occurrence.

Each piece of linguistic evidence generally supports the others and the stylistic evidence in the play. My counts for *ye, you, doth, does, hath, has, them* and *'em* in *TNK* (with contractions expanded) are given in Table 7–1 on page 297.

In describing Fletcher's normal rate of use of the two auxiliary verb forms, Hoy states that *doth* occurs at most three times in a single play; *hath* at most 6 times. However, my counts in *Demetrius and Enanthe* (the manuscript version of *The Humorous Lieutenant* used in this study) show that *doth* occurs 10 times and *hath* 23 times. In his series of articles in *Studies in Bibliography,* Hoy relies on the Folio edition of this play. He makes no mention of the discrepancies for these counts between the manuscript and Folio versions in either this series or in the introduction to his edition of the play in Bowers' series [18]. (The two versions of the text were discussed earlier in Section 2.4.1, which begins on page 43.)

He does note (in both places) that the scribe of the manuscript, Crane, is fairly faithful in reproducing Fletcher's preference for *ye*. Bald (in *Bibliographical Studies in the Beaumont and Fletcher Folio of 1647* [6]) finds that Crane uses

The Two Noble Kinsmen

	Words	Attr.*	*does/doth*		*has/hath*		*'em/them*		*ye/you*	
I.i	1821	Sh	2	1	0	2	2	5	0	36
I.ii	954	Sh	1	0	0	3	1	2	0	6
I.iii	804	Sh	0	0	3	1	1	2	0	10
I.iv	413	Sh	1	0	0	1	5	4	0	2
I.v	108	Sh	0	0	0	0	0	0	0	1
II.i	497	Sh	0	0	0	0	3	3	0	6
II.ii	2402	Fl	0	0	2	1	10	0	7	31
II.iii	744	Fl	0	0	4	0	1	0	2	7
II.iv	288	Fl	0	0	2	0	0	0	0	0
II.v	573	Fl	0	0	0	0	0	0	0	38
II.vi	355	Fl	0	0	1	0	0	0	0	1
III.i	1051	Sh	0	0	1	1	0	2	1	25
III.ii	343	Sh	0	0	2	1	0	1	0	0
III.iii	502	Fl	0	0	1	0	0	1	0	20
III.iv	250	Fl	0	0	1	0	0	0	0	2
III.v	1241	Fl	1	0	0	0	1	0	8	19
III.vi	2717	Fl	0	0	4	0	10	1	11	60
IV.i	1353	Fl	0	0	3	0	4	2	3	26
IV.ii	1349	Fl	0	0	8	0	11	3	0	8
IV.iii	877	Fl	0	0	1	2	0	1	0	10
V.i	1392	Sh†	0	0	1	0	2	2	3	9
V.ii	1039	Fl	0	0	6	0	0	0	4	40
V.iii	1211	Sh	0	0	2	0	2	4	0	16
V.iv	1158	Sh	0	0	1	4	2	0	0	18
Pro.	273	?	0	0	1	0	0	0	0	3
Epi.	169	?	0	0	1	0	0	0	8	0

*Attribution by Hoy [55].
†Except for the first 33 lines (276 words).

Table 7–1: Counts in *TNK* of some features studied by Hoy

hath nine times where the Folio uses *has; doth* occurs seven times where F1 reads *does*. He notes that Crane showed a similar preference for the older forms in his transcription of Middleton's *A Game at Chess*. Hoy does not appear to address this question concerning *doth* and *hath* as far as I can discover. It appears that the occurrences of *hath* and *doth* may be due to Crane; in any case, this is a good example of how counts for such forms can be affected by a scribe.

Like many of his predecessors, Hoy does not question the division in any great detail. In comparing works of known authorship he does not examine samples smaller than a play. The features he counts do not usually occur often enough to allow the examination of individual scenes independently. The number of occurrences in most of the smaller scenes of *TNK* is so small that one cannot really claim that Hoy's results assign them to either author. One can claim only that, overall, the counts are not inconsistent with the accepted view. No single scene in *TNK* has counts for these features that might contradict the accepted division (with the possible exception of IV.iii).

7.1.2 Discriminant Analysis Results

The 1634 Quarto text of *The Two Noble Kinsmen* was divided into scenes, and counts were made of the function word markers. (Proudfoot's edition [38] is the source of the scene divisions and the line numbers that are quoted in the following discussion.) Program KERCON was used to determine the posterior probabilities of authorship $P(\omega_i \mid \mathbf{x})$ using the three sets of words (T1, T2 and FR1) evaluated in Chapter 6. Table 7–2 on page 299 presents these results. For each scene the table lists: (1) the number of words (after contractions have been expanded); (2) the generally accepted attribution (according to Proudfoot and Hoy); (3) the classification result and the posterior probability for each of the three word-sets; and, (4) the "verdict" of these three results.

This last column indicates whether the scene is allocated to either author or whether the scene cannot be classified because the posterior probabilities are too

The Two Noble Kinsmen

			Posterior Probabilities							
	Words	Attr.*	Set T1		Set T2		Set FR1		Verdict
I.i	1821	Sh	Sh	1.000	Sh	0.968	Sh	0.988	Sh
I.ii	954	Sh	Sh	1.000	Sh	1.000	Sh	1.000	Sh
I.iii	804	Sh	Sh	0.998	Sh	0.988	Sh	1.000	Sh
I.iv	413	Sh	Sh	1.000	Sh	1.000	Sh	1.000	-short-
I.v	108	Sh?	Fl	0.820	Sh	0.965	Fl	0.880	-short-
II.i	497	?	Sh	1.000	Sh	1.000	Sh	0.999	Sh
II.ii	2402	Fl	Fl	0.911	Fl	0.755	Fl	0.853	Fl
II.iii	744	Fl	Sh	0.644	Sh	0.836	Sh	0.921	Sh
II.iv	288	Fl	Fl	0.653	Fl	0.999	Sh	0.899	-short-
II.v	573	Fl	Fl	0.999	Fl	1.000	Fl	1.000	Fl
II.vi	355	Fl	Fl	0.997	Fl	1.000	Fl	0.993	-short-
III.i	1051	Sh	Sh	1.000	Sh	0.972	Sh	1.000	Sh
III.ii	343	Sh?	Sh	0.932	Sh	0.974	Sh	0.932	-short-
III.iii	502	Fl	Fl	0.995	Fl	0.987	Fl	0.982	Fl
III.iv	250	Fl	?†	—	Sh	0.818	Fl	1.000	-short-
III.v	1241	Fl	Sh	0.603	Sh	0.963	Fl	0.523	?
III.vi	2717	Fl	Fl	0.950	Fl	0.982	Fl	0.586	Fl
IV.i	1353	Fl	Sh	0.853	Sh	0.678	Fl	0.703	?
IV.ii	1349	Fl?	Sh	0.619	Fl	0.958	Fl	0.790	?
IV.iii	877	Fl?	Sh	1.000	Sh	1.000	Sh	1.000	Sh
V.i	1392	Sh‡	Sh	1.000	Sh	0.998	Sh	0.994	Sh
V.ii	1039	Fl	Sh	0.526	Fl	0.526	Fl	0.810	?
V.iii	1211	Sh	Sh	0.998	Sh	0.972	Sh	0.927	Sh
V.iv	1158	Sh	Sh	0.999	Sh	0.995	Sh	0.952	Sh
Pro.	169	?	Fl	1.000	Fl	1.000	Fl	0.999	-short-
Epi.	273	?	Fl	0.796	Sh	0.948	Fl	0.741	-short-

*Attribution of each scene according to Proudfoot [38] and Hoy [55].
†Maximum likelihood procedure failed to estimate smoothing parameters properly.
‡Except for the first 33 lines (276 words).

Table 7–2: Classification results for *The Two Noble Kinsmen*

low or disagree. (A question mark indicates that the scene remains unclassified.) In reaching this decision, the criteria discussed in Section 6.5.5 have been used: using a reject threshold of 0.20, scenes are accepted for classification if all three probabilities are not rejected and agree on an author, or if the probability for only one word set is rejected but the classification still agrees with the other two sets. (This resulted in a successful classification rate of 96.5% for the design set and 87.5% for the test set, where 2 of the 88 scenes (2.3%) were misclassified.)

Seventeen of the the 26 scenes are longer than 500 words, and another, II.i, is so close (497 words) that the rules have been bent, and it is included with the longer scenes. (To satisfy the curious, the probabilities for the shorter scenes have been listed, but no conclusions should be drawn regarding these. The verdict column for these scenes contains the text "-short-".) All but 4 of these 17 scenes are classified, and for two of these scenes, II.iii and IV.iii, the decision disagrees with Proudfoot and Hoy's assignment. In the following paragraphs, the words rates in each scene will be examined to try to learn what particular words might be responsible for the assignments listed in Table 7–2. As in the examination of rejected and misclassified scenes in the last chapter, the rate of occurrence of each word will be compared to the Shakespeare and Fletcher mean rate using z-scores, calculated for each author according to (6.12) on page 282.

Acts I and V

It is very encouraging that the classification procedure credits Shakespeare with the first three scenes of Act I and with scenes i, iii and iv of Act V. Critics have always agreed that these scenes, with their highly complex verse, are his, with one short exception. The first 33 lines (276 words) of V.i (up to the exit of Palamon and his knights) resemble Fletcher's style of verse, and many (Hoy, for example) view this passage as an interpolation by the younger playwright. If this is true then its inclusion does not affect the result based on these markers; the entire scene is allocated to Shakespeare. The passage is too small to be analyzed independently. If one accepts Fletcher's authorship of these lines, one

must then conclude that interpolations of this length occurring in long scenes by Shakespeare may not affect the classification results dramatically.

Examination of the z-scores for the word rates in these Shakespearean scenes indicates that several of the markers occur at very different rates than normally found in Fletcher's work. The results appear primarily due to the number of occurrences of the more common markers *in, of* and *the*. In particular, the first three scenes of the first act are all marked by very high occurrences of *in* ($z_{\mathrm{Sh}} \geq 2$, $z_{\mathrm{Fl}} \geq 4.5$). All six scenes are also marked by high rates of *which*, a Shakespeare marker only included in set T1. (While the short length of I.iv is outside the range accepted for examination, the general pattern of the rates described for the other scenes of Act I is also evident in this scene.)

The one scene in these two acts that is usually accepted as Fletcher's is V.ii. This scene is not classified since the posterior probabilities for set T1 (0.526 for Shakespeare) and set T2 (0.526 for Fletcher) are rejected. Examination of the z-scores for the variables shows that no particular word occurs at a very non-Fletcherian rate. In fact, two variables favor the accepted position: for *too* $z_{\mathrm{Sh}} = 2.2$, and for "Infreq-Sh+" $z_{\mathrm{Sh}} = -2.1$ (only 1 occurrence). As noted in the last chapter, z-scores are sometimes useful for indicating that a particular word is greatly affecting a result. For V.ii they reveal nothing, and one must conclude that, in this instance, the marker words do not supply enough evidence to assign the scene to one author or the other.

Act II

The results for the first two scenes of Act II agree with the accepted attributions. The first scene looks very much like Shakespeare. A very high rate of *of* contributes to the result; 20 occurrences result in a rate of 40.2: $z_{\mathrm{Sh}} = 2.9$, $z_{\mathrm{Fl}} = 6.0$. Examination of the text reveals nothing that would explain this unusually high rate. The rates for *the, in* and *too* also favor the elder dramatist. A high rate for one Fletcher marker that occurs sets T1 and FR1, *all* ($z_{\mathrm{Sh}} = 2.8$), is overwhelmed by the other evidence.

The second scene has a rejected posterior probability for set T2, but since the assignment agrees with that of the other two non-rejected probabilities, the scene is assigned to Fletcher. Almost none of the word rates are very different from either writer's averages. The only z-score in all three word sets that is greater than 2.0 is that for *must*. The 17 occurrences in this scene (which result in a z-score for Shakespeare of 2.9) are often found in clusters in the text (lines 22, 27 and 28, lines 45 and 48, lines 203 and 208, lines 224 and 226, lines 271, 273 and 276).

The results for II.iii is the first that disagrees with the accepted evidence. It begins with Arcite lamenting his banishment; he is interrupted by four rustics who tell him of the games at the court (after some slightly bawdy exchanges). Arcite then resolves to disguise himself, return to the court and compete. Parts of the scene are usually treated as verse by modern editors, although the quarto prints them as prose. Some have felt that this distinction bears on the authorship question (see Proudfoot's discussion of Bertram [38, p. xxv], and Littledale's remarks about Fleay [37, Vol. 1, page 136]). Littledale notes that the scene is "of course, by Fletcher," and quotes Spaulding's judgement that "neither this scene, nor the following, ... have anything in them of particular notice." The metrical evidence presented by Littledale in his second volume and Oras' counts of extra monosyllables [119] certainly support the case for Fletcher, but these results might be questioned (that is, more so than usual for metrical tests) because of the rough texture of the verse (if verse it is).

As inclined as one might be to agree with Spaulding, the occurrence of function words in II.iii more closely resembles scenes known to be Shakespeare's than Fletcher's. The probability resulting from set T1 is below the reject threshold (0.644) but agrees with the other two non-rejected probabilities. These values (0.836 and 0.921) are not as high as the probabilities for the Shakespearean scenes of Acts I and V, so perhaps the evidence is less strong. Examination of the z-scores indicates a high rate for the Shakespeare marker *the*. 30 occurrences result in rate of 40.3, $z_{\mathrm{F1}} = 2.3$. The rate for *in* is approximately equal to the

Shakespeare average ($z_{\mathrm{Fl}} = 1.5$). On the other hand, rates for *must* and *all* favor Fletcher's claim ($z_{\mathrm{Sh}} = 2.0$ and 1.6, respectively). There are low rates for both sets of the pooled infrequent markers included in set T2.

The seems to be the crucial variable here. Only 5 Fletcher scenes of the 106 in the design set have higher rates for the definite article. Examination of a concordance reveals nothing that might attribute any occurrences to a stylistic device or repetition, although a slightly higher proportion of the occurrences fall in the rustics' exchanges. It is of course possible that both writers had a hand in the scene, although there is no positive evidence to suggest this. Of the forms examined by Hoy (Table 7–1), the pair *ye* (2 occurrences) and *you* (7 occurrences) supports Fletcher's case. No scene in the Shakespeare design or test set has a comparable proportion of occurrences of these forms.

All three results assign II.v to Fletcher. It contains an unusually high number of occurrences of *are* (13 occurrences), which is only included in set FR1. This is untypical of both writers, $z_{\mathrm{Fl}} = 4.6$ and $z_{\mathrm{Sh}} = 6.7$, and seems to be due to the plot. In this scene Theseus and his followers interview the disguised Arcite, with a lengthy discussion of identity and character ("are you a gentleman," "are you his heir," "you are perfect," "you are mine," "you are hers," "you are a noble giver," "you are a horseman"). A high rate for *all* also points to Fletcher, and this assignment is strengthened by two occurrences of *dare* ($z_{\mathrm{Sh}} = 5.8$).

Act III

Act III presents no major surprises. The posterior probabilities for III.i are strongly in Shakespeare's favor; the z-scores show that the rates of *in* and *of* are unlike the Fletcher averages. Program KERCON and literary scholars also agree that III.iii is Fletcher's. It contains high rates for *too* and *now* and a low rate for *of*. The last scene in the act, III.vi, is also assigned to Fletcher, although one probability is rejected. For set FR1 (which is the least accurate of the three sets) $P(\mathbf{x} \mid \omega_i = \mathrm{Fl}) = 0.59$. The z-scores are most unhelpful in explaining this

result. Fletcher still gets credit because of the the other two probabilities, which are due in part to 5 occurrences of *dare* ($z_{\mathrm{Sh}} = 2.8$).

III.v remains unclassified. This scene contains speeches by the schoolmaster Gerrold and the presentation of the morris dance (whose participants appear in Beaumont's masque) before Theseus and his court. These features instantly recall the mechanical's performance before the same monarch in *A Midsummer Night's Dream*. Hickson (mistakenly believing that the play was written some four years earlier) describes this as the "imitation of a young and experienced writer" [48, p. 57*]. The scene again contains a much larger proportion of *ye* to *you* (11 to 60) than found in scenes known to be by Shakespeare.

Like II.iii this scene contains more occurrences of *the* than one expects to find in Fletcher's work. (In fact, the rates for the two scenes are equal; $z_{\mathrm{Fl}} = 2.3$). The z-scores for the other variables are not informative, except "Infreq-Fl+;" a low rate for this group of words appears to combine with *the* to produce the one non-rejected probability: set T2's value of 0.96 for Shakespeare. The rates for a number of the marker words are low for either author. The large proportion of *ye* to *you* (8 to 19) in the scene clearly supports the accepted view. Some have conjectured that Beaumont assisted in this scene, basing this claim only on the presence of his morris dancers. Discriminant analysis can tell us nothing more than this: the function word evidence does not unambiguously point to either Shakespeare or Fletcher.

Act IV

The method also fails to allocate the first two scenes of Act IV. These are usually attributed to Fletcher. Proudfoot notes that some doubts have been expressed about IV.ii, but Hickson judged it to be "Fletcher's masterpiece." The proportion of *ye* to *you* is not so large here; the counts in IV.ii (3 and 26) are close to the counts in II.iii of Shakespeare's *Twelfth Night* (3 and 27). *TNK* IV.i–ii contain more occurrences of *'em* than normally found in Shakespearean scenes,

although *The Tempest* III.ii has an unusually high proportion (5 occurrences of *'em*, 3 of *them*).

IV.i is another scene without strong indications for any of the marker words. None of the z-scores for either author are much greater than 1.0 or less than -1.0, with the exception of *must*. This Fletcher marker is only included in set FR1 (which produces a probability of 0.70 for Fletcher), and the 9 occurrences result in a somewhat high rate (6.7, $z_{\mathrm{Fl}} = 1.5$, $z_{\mathrm{Sh}} = 2.7$). The second scene of Act IV also contains a relatively large number of occurrences of *must* ($z_{\mathrm{Sh}} = 2.3$). As in II.ii the occurrences of *must* in these two scenes in Act IV fall into clusters of 2 or 3 within a span of about 4 or 5 lines.

Unlike its predecessor, IV.ii contains strong but conflicting evidence for several markers. A large number of words in the pooled set of Fletcher markers (such as *yet, nor* and *still*) produce a Fletcher-like rate for this variable; the Shakespeare z-score is 2.8 and the probability for set T2 is 0.96 for Fletcher. However, the rate of occurrence for *of* is high for Fletcher ($z = 2.3$), and this seems to cause the outcomes for the other two word sets to be ambiguous.

The last result to be discussed is the most interesting finding in this scene by scene analysis of the play. The authorship of IV.iii was debated at length by the Victorians. Littledale pronounced: "On the way in which we determine the authorship of this scene, must depend our view of Shakespeare's share in the play as a whole" [37, Vol. 1, p. 155]. Everyone was quick to attribute the play's opening and closing scenes of high rhetoric to Shakespeare. The development of the sub-plot caused more concern, since many Victorians were unwilling to accept that Shakespeare had anything to do with the bawdiness of the rustics' speeches, the sexual aspects of the jailer's daughter's mad ravings or the bed-trick finally used to cure her.[4]

[4]The degree of these objections is often entertaining to the modern mind. In Littledale's edition and in the *Transactions of the New Shakspere Society,* one encounters repeated references to the "trash" of the sub-plot. One wonders how they interpreted the bawdy scenes in accepted masterpieces, such as *Lear* and *Hamlet.* In Hickson's 1847

Since IV.iii is in prose, no easy answer could be supplied by metrical tests. Hickson discusses the similarities between the first meeting of the doctor and the daughter in IV.iii and the scenes in Shakespeare where physicians deal with madness (such as in *King Lear* and *Macbeth*) [48, p. 50]. Most critics have regarded these as Fletcher's imitation of Shakespeare, but Hickson argues that the similarities in language and in the gradations from a "mind diseased" to full madness do have the air of being original. This would indicate that Shakespeare did not fully abandon the jailer's daughter to his collaborator after introducing her in II.i, and leads one to make interesting comparisons of her progression through stages of madness to Ophelia, Lear and Lady Macbeth. Many modern editors seem to place little emphasis on such comparisons, accepting the scene as Fletcher's and the similarities as imitation.

In Section 5.3 the occurrence of *thereto* in this scene was noted. Shakespeare uses words beginning with *where-* or *there-* almost 12 times as often as Fletcher. This word occurs 12 times in the 20 Shakespeare control-set plays but does not occur in any of the Fletcher texts examined in this study. Hoy's evidence does not support an assignment to Fletcher; only 9 scenes of the 106 in the Fletcher design set contain two or more occurrences of *hath*. Scenes with 10 or more occurrences of *you* without an occurrence of *ye* cannot be found among the design-set scenes (although *The Women's Prize* II.i has 1 *ye* for 10 occurrences of *you*).

The evidence provided by the analysis of function words strongly supports Shakespeare's authorship of IV.iii. The posterior probabilities for all three sets of words are as high or higher than any scene in the Shakespearean portion of the first and last acts. Like those scenes, IV.iii contains relatively large numbers of

---

article, while asserting Shakespeare's authorship of *TNK* IV.iii, he criticizes Knight for his purge of the sub-plot in an earlier edition. To support his contention he quotes a speech, in which the daughter imagines hell, beginning: "Lords and Courtiers, that haue got maids with Child, they are in this place. . . ." When this article was reprinted in the 1874 *NSS Transactions,* this passage is printed with blank space for the clause "that haue got maids with Child." A footnote explains: "In the original a qualifying phrase here occurs, very shocking to Mr. Knight."

in ($z_{\mathrm{Fl}} = 3.9$, $z_{\mathrm{Sh}} = 1.7$), *of* ($z_{\mathrm{Fl}} = 2.7$) and *which* ($z_{\mathrm{Fl}} = 2.5$). These rates for *in* (21.7) and *of* (25.1) are very different from observed values in Fletcher's known works. Of the 106 scenes in the design set, the highest rate of *in* is 18.6; only 5 scenes have rates above 15.0. This contrasts sharply to the 265 Shakespeare scenes, where 14 scenes have rates higher than *TNK* IV.iii (96 scenes have rates higher than 15.0). For *of*, only one Fletcher scene has a rate as high as this scene, compared to 53 scenes in the Shakespeare design set.

7.1.3 Summary

A reader who accepts the traditional division of *The Two Noble Kinsmen* should be encouraged by the results presented in this section. The function word rates in all the scenes attributed to Shakespeare are extremely similar to the 265 Shakespeare scenes in the design set: the posterior probabilities calculated by the kernel classification program are all above 95%.

However, an observant reader will note that only two scenes are attributed to Fletcher with such high probability: II.v and III.iii. In fact, the four scenes that could not be classified by the procedure are all usually attributed to Fletcher. The evidence put forward by Hoy supports the orthodox attribution in these cases. However, the case for Shakespeare's authorship of IV.iii is strong; the counts of *hath* and *you* and the single occurrence of *thereunto* lend support to the high probabilities resulting from the discriminant analysis based on function words. No such support can be offered for the method's assignment of II.iii to Shakespeare, and few critics are likely to accept the conclusion that this scene is his.

7.2 *Henry VIII*

The Famous History of the Life of King Henry the Eighth was first published in the 1623 Shakespeare Folio. The text is quite a good one, unusual for its elaborate stage directions, which are needed to convey the pageantry of the play. Like other plays written late in Shakespeare's career, the language is often very complex, but the text shows no signs of corruption. While extensive stage directions often indicate prompt copy, some features of these (such as descriptions of gestures and emotions) suggest instead authorial intent. The speech prefixes are consistent for the most part, and entrances and exits are clearly marked. Thus scholars agree the source behind the Folio text was probably a fair copy written out by a scribe rather than foul papers. If indeed two authors were responsible for the play, the transcript must have been carefully edited in preparation of the fair copy. Humphreys believes that the manuscript was prepared for reading [129].

Episodic rather than dramatic in structure, the play in turn depicts the rise and fall of the fortunes of Buckingham, Katherine and Wolsey. The author or authors have borrowed heavily from two Elizabethan chronicle histories, Holinshed's *Chronicles* and Foxe's *Acts and Monuments* (the latter mainly for the story of Cranmer in Act V), and at some points the text is little more than the source passages turned into verse. (A number of scholars have attempted to make judgements regarding authorship based on the different ways the sources are used. The results of some of these studies are outlined in *Evidence and Authorship* [33, pp. 457–478].)

The external evidence for dating the play is unusually plentiful, for on 29 June 1613 the famous Globe theater, home of Shakespeare's company since 1599, burned during an early performance of the play. A number of private letters document this event; several give the title of the play as *All Is True*. The general descriptions of the performance found in these letters and the emphasis that the

prologue places on "truth" lead most critics to accept the identity of this play with *Henry VIII,* despite the fact some details of the performance provided in one letter do not quite fit the text. The fire started when cannons were fired off to mark the king's entry at a masque at Wolsey's house (Act I Scene iv). The thatch in the roof caught, and the structure burned to the ground. Sir Henry Wotton's letter notes that no one died, although one man's breeches were set on fire: "That would have perhaps broiled him, if he had not by the benefit of a provident wit put it out with bottle ale."[5] The play was probably written because of the wedding of the Princess Elizabeth to the Elector Palatine in February 1613, although it is not clear if it was part of the official celebrations. In any case, the public attention surrounding this state occasion could help explain the play's unusual attention to pageantry and Shakespeare's return to the history play genre.

7.2.1 Past Studies of Internal Evidence

The first serious attempt to credit Fletcher with a share of the composition of *H8* was Spedding's 1850 article "Who Wrote Shakespeare's *Henry VIII?*"[6] Like many others before and since, Spedding saw in the play an incoherent design. He regards the characterization of the king as weak and cannot understand why the play follows the sympathetically-treated downfalls of Buckingham and Katherine with a celebration of Henry and Anne's success in producing an heir. Thus the play as a whole was "weak and disappointing." Starting from a remark by Tennyson that some of the passages resembled Fletcher's verse, Spedding divided the play scene by scene between Shakespeare and Fletcher on the basis of "the general effect produced on the mind, the ear, and the feelings by a free and broad perusal" [152, p. 7*]. (Spedding divided one scene, III.ii, between the two

---

[5]Reprinted in most modern editions, for example Humphreys' [129].

[6]This article was reprinted with the title "On the Several Shares of Shakspere and Fletcher in the Play of *Henry VIII*" by the New Shakspere Society in their 1874 *Transactions* [152]. All my references are taken from this version of the article.

writers, crediting Shakespeare with the first part and Fletcher with the second. These two parts will be identified as "III.iia" and "III.iib.")

Hickson immediately announced that he had come to roughly the same conclusions independently. After some discussion between them, slight adjustments were made to the attribution of scenes (mainly confirming Fletcher's responsibility for Act IV, of which Spedding was initially less sure). This result became the orthodox division of the play, and is still reported as such in modern editions (despite new conclusions by Hoy, founded on a more objective analysis of linguistic evidence).

At the end of the article, Spedding adds a table that supports his division with measurements of the proportion of lines with feminine endings. Alexander, attacking the interpretation of metrical and stylistic evidence by the supporters of collaboration in 1939, notes that he does not pursue the objective analysis of metrical statistics very far [1]. The degree of variability in Shakespeare's known works is never considered; in fact, statistics are not presented for any other play by either author. But other scholars followed Spedding's lead, and many metrical tests were subsequently applied to *H8*, *TNK* and other works by Shakespeare. Such tests were used to support not only Spedding's division but the general case for the disintegration of Shakespeare's canon.

Those supporting Shakespeare's sole authorship have always placed considerable importance on the inclusion of the play in the 1623 Folio. Shakespeare's part-authorship of *TNK* can be viewed as evidence that Heminges and Condell would not have included a collaborative play. As noted earlier, one can conjecture that this exclusion was due to unavailability of a text at the time. Part of the reaction against the wilder excesses of the disintegrators has been an acceptance of Heminges and Condell's claim that they present Shakespeare's texts "absolute in their numbers, as he conceived them" [134, p. 7]. But as Mincoff points out, Heminges and Condell's authority was accepted only after being vindicated by the same "philological method" that questions the single-author position for *H8* [105]. Law notes that blind acceptance of their claim would credit Shakespeare

with the Hecate scenes in *Macbeth* [67] (not to mention *Timon*, *Titus Andronicus* and *Pericles*, whose authenticity is questioned by many).

The history of the discussions surrounding *H8* parallels the study of *TNK*, and naturally many of the studies reviewed in the previous section also consider this play. One contrast between the two questions is important; in *H8* many of the best-loved speeches in the play are given to Fletcher by Spedding and Hickson's division. These include Buckingham's farewell (II.i), Wolsey's fall (III.iib) and Katherine's final speeches (IV.ii). This is one of the main reasons that assertions of Fletcher's participation in *H8* have been more controversial than those of Shakespeare's unspectacular role in *TNK*. Massinger was also advanced by those who wished to deny Shakespeare's participation in *H8*, mainly on the basis of parallels between the play and Massinger's known works. But several critics have shown that Massinger was a frequent borrower from the elder dramatist's plays. As in the case of *TNK*, Farnham's analysis of all three writers' use of contractions [35] shows that the use of such forms in *H8* is extremely unlike Massinger but conforms to the other two dramatists' habits.

A very important study published in 1947 by Partridge turned attention back to linguistic differences between the two playwrights, after several decades of neglect. In *The Problem of "Henry VIII" Reopened*[7] Partridge examines the older writer's fondness for auxiliary *do* where no emphasis is intended; such "expletive" occurrences are often used for metrical purposes. (The use of auxiliary *do* in Early Modern English is described by Barber [7, pp. 263–267].) This usage occurs 45 times in the scenes attributed to Shakespeare and only 5 times in Fletcher's portion. In demonstrating that Fletcher rarely uses *do* in this way, Partridge quotes figures from *Bonduca* and *The Faithful Shepherdess* but does not present a detailed analysis in a large number of plays. Nor does he examine the degree of internal variation of this feature in other Shakespeare plays.

[7]This monograph was revised and re-published as part of *Orthography in Shakespeare and Elizabethan Drama* [120], and my discussion is based on this publication.

Partridge also examines other reflections of Shakespeare's preference for older linguistic practices (later analyzed by Hoy): his use of *doth* and *hath* and his reluctance to use *ye* and *'em*. He also reviews Farnham's study of contractions. Partridge states that one cannot explain certain "exceptional uses" of these markers in parts of the play where they should not occur (according to the accepted division). He regards all of these traits as "preponderant" rather than "exclusive" and believes that Fletcher completed an unfinished draft of Shakespeare's. (This theory has not met with great critical acceptance.)

The most important attack against Fletcher's participation since Alexander's was made by Foakes in the introduction to the 1957 Arden edition [130]. While recognizing that some internal evidence resembles Fletcher's practice, he argues that the evidence is not strong enough to overturn the evidence for sole authorship. These include its inclusion in the 1623 Folio, the use of sources, imagery and the "compassionate tone and outlook" that *H8* shares with Shakespeare's romances.

Foakes also stresses the alterations that scribes and compositors could make in forms such as *doth, hath, ye, 'em* and other contractions. He notes that the number of occurrences of *ye* in the part attributed to Fletcher is much lower than found in his play *Bonduca*. Farnham's evidence for contractions "is so narrow as to establish little more than that both authors were inconsistent in their usage" (p. xix). He also notes that Shakespeare's *Cymbeline* can be divided to produce a distribution of some contractions (*i'th, o'th, to'th* and *by'th*) similar to that found in *H8*.

Foakes accurately notes that "support for Fletcher has nearly always been associated with condemnation of *Henry VIII* as bad or lacking unity, and belief in Shakespeare's authorship with approval of the play" (p. xxii). His belief in a single author may not have been as strong as they come across (he later recants somewhat in his preface to the second edition), since at one point he comments that, "if Fletcher has to be introduced," then he must have worked as an occasional reviser who contributed one or two scenes. Foakes states that he can

by no means credit Fletcher with the unaided composition of II.ii, IV.i or V.iii. He also notes an important characteristic of the occurrences of *ye* in the text. In a number of instances, these occur clustered together in three or four lines. Hoy places great significance on this in his analysis of the play in the final part of his series in *Studies in Bibliography* [55].

After demonstrating his procedures with *TNK*, Hoy begins his examination of *Henry VIII* with a reply to Foakes' criticism of the evidence for collaboration. He turns Foakes' fears of transmission alterations against him. Foakes' objection that the use of *ye* in *H8* is not Fletcherian enough is explained by findings which show that one of the compositors who set the Folio text frequently altered *ye* to *you*. (This result, due to Williams, is discussed by Hoy and in *Evidence for Authorship* [34, p. 474].) But Hoy does accept that the local clusters of *ye* in II.i-ii, III.iib and Act IV are signs of "Fletcher the interpolator not Fletcher the original author." Two of these scenes occur in pages set by Compositor X, who appears not to have tampered with *ye* when setting his copy, and Hoy supplements his contention for the other scenes with stylistic evidence. (My counts for *ye, you, doth, does, hath, has, them* and *'em* in *H8* are listed in Table 7–3 on page 314.)

Thus, Hoy feels that I.iii–iv, III.i and V.ii-iv are wholly Fletcher's, and he highlights examples of the younger dramatist's syntactic and rhetorical characteristics in these scenes. Several of these are quite striking (in particular, the parenthetical inversions described by Hoy in item (c) on page 82); some are less so. After his detailed examination of the entire Beaumont and Fletcher canon, Hoy's knowledge of Fletcher's style cannot be disputed. Overall his evidence and analysis are very convincing, although it does demonstrate that the most distinctive linguistic features in the text may reflect compositorial interference or one author's revision of the other's work.

Mincoff's spirited article "*Henry VIII* and Fletcher" [105] is perhaps the most comprehensive reply to Foakes' arguments for sole authorship. He reviews most aspects of the evidence, from feminine endings to the use of *ye*, concluding

Henry VIII

	Words	Attr.*	*does/doth*		*has/hath*		*'em/them*		*ye/you*	
I.i	1868	Sh	1	0	3	5	2	5	1	26
I.ii	1742	Sh	0	1	1	3	2	5	0	24
I.iii	587	Fl	0	0	1	0	7	0	0	3
I.iv	941	Fl	0	0	1	0	12	1	4	27
II.i	1439	both	1	0	1	0	4	0	4	20
II.ii	1220	both	1	0	1	1	2	1	3	12
II.iii	898	Sh	0	0	0	1	0	0	1	25
II.iv	1924	Sh	1	0	0	3	0	2	0	54
III.i	1525	Fl	0	0	2	0	5	0	20	30
III.iia	1663	Sh	4	1	1	7	1	3	0	38
III.iib	2185	both	1	0	1	1	2	1	6	37
IV.i	999	both	0	0	0	0	3	1	3	14
IV.ii	1431	both	1	0	0	0	3	1	5	20
V.i	1507	Sh	1	0	0	3	0	4	0	44
V.ii	296	Fl	0	0	0	0	3	0	0	3
V.iii	1550	Fl	0	0	0	0	3	0	12	40
V.iv	807	Fl	0	0	0	0	13	0	8	15
V.v	653	Fl	1	0	1	0	1	0	7	3
Pro.	268	?	0	0	0	0	0	1	2	4
Epi.	132	?	0	0	0	0	2	0	0	0

*Attribution by Hoy [55].

Table 7–3: Counts in *H8* of some features studied by Hoy

that the number of tests supporting collaboration outweigh the ill-supported objections for sole-authorship. In an interesting examination of the early metrical studies, Mincoff notes that Fletcher's rates for feminine endings and end-stopped lines differ between his unaided plays and parts of the plays he wrote with Beaumont. He appears to adapt his style to Beaumont but not to some of his other collaborators (like Field and Massinger). Since these same characteristics of his prosody are reduced in his part of *TNK* and *H8*, Fletcher must have also adapted his style to Shakespeare. Mincoff maintains that this answers the charge that the scenes attributed to Fletcher in *H8* are not enough like Fletcher's other work.

He dismisses as "merely frivolous" Foakes' concern about alterations by scribes, editors or compositors. Although he does not argue from a bibliographic standpoint, he doubts that the occurrences of *ye* could have been added by coincidence only to scenes marked by Fletcher's style and meter. In a more valuable criticism, he maintains that the imagery which Spurgeon [157] and Foakes use to support Shakespeare's sole authorship is frequent in Elizabethan writings. In fact, many of the "Shakespearean" images that run through the play are found in Fletcher's *Valentinian.*

Citing I.iii as "the most unmistakably Fletcherian scene in the whole play" (page 248), Mincoff uses this short scene to illustrate many of younger dramatist's characteristics. Not only is the number of feminine endings quite high in this scene (7, compared to 3 in I.i–ii, which contain 7 times as many lines), they affect the flow of the verse in a different manner from feminine endings in Shakespeare. (The particular characteristics of Shakespeare's and Fletcher's use of double-endings have also been examined by Oras [119] and Law [67].) Mincoff also notes a number of the more striking syntactic parallels between Fletcher's part of *H8* and his unaided work. He also identifies Fletcher's typical humor: the joining together of incongruities pointed by alliteration. Examples cited in *H8* include lines 23–35 of I.iii: "fool and feather," "fights and fireworks," "tennis and tall stockings."

Later in his paper he does find it odd that this scene has a very high proportion of run-on lines, a Shakespearean characteristic, but also contains the highest proportion of feminine endings.[8] At several points one wonders about Mincoff's priorities: does he accept the objective analysis of linguistic features only when it supports his subjective stylistic judgements? For example:

> The crowd scene, V.iv., bears his [Fletcher's] mark very unmistakably, more so perhaps than Cranmer's trial [V.iii.], although the metrical figures are not all reminiscent of him and the larger part is in prose, which he avoided as a rule.

Nevertheless, Mincoff's article is a comprehensive examination of different types of evidence. His examinations of imagery in other texts and changes in Fletcher's traits in other collaborations are certainly noteworthy.

Since the outburst of articles provoked by Foakes' edition of the play, most editors preparing modern editions have accepted the dual authorship conclusion. These include Maxwell [131], Humphreys [129], Schoenbaum [127] and the editors of the Riverside edition (who also include *TNK* in their edition of Shakespeare's works). A notable exception is Bevington [126], who falls back on the authority of Heminges and Condell. Some other critics have continued to maintain the sole authorship position, including Knight [64] and Sprague, who argues interestingly from the point of view of theatrical production [156].

7.2.2 Discriminant Analysis Results

The Folio text of *Henry VIII* was divided into scenes according to the division of most modern editors, which only differs from the Folio scene division by the introduction of a new scene division after line 35 in V.ii. In addition, the second scene of Act III was divided in two according to Spedding's division, and

[8]The scene is quite short, containing only 67 lines of verse. Throughout the history of metrical tests, little attention has been paid to determining the possible variation in short samples of an author's undisputed work.

Henry VIII

		Attrib. by*		Posterior Probabilities							
	Words	Sped.	Hoy	Set T1		Set T2		Set FR1		Verdict
I.i	1868	Sh	Sh	Sh	1.000	Sh	1.000	Sh	1.000	Sh
I.ii	1742	Sh	Sh	Sh	1.000	Sh	1.000	Sh	0.998	Sh
I.iii	587	Fl	Fl	Sh	0.985	Sh	0.729	Sh	0.802	Sh
I.iv	941	Fl	Fl	Fl	0.790	Fl	0.896	Fl	0.991	Fl
II.i	1439	Fl	both	Fl	0.932	Sh	0.737	Fl	0.560	?
II.ii	1220	Fl	both	Fl	0.955	Sh	0.805	Sh	0.717	?
II.iii	898	Sh	Sh	Sh	0.902	Fl	0.732	Fl	0.700	?
II.iv	1924	Sh	Sh	Sh	1.000	Sh	0.998	Sh	0.997	Sh
III.i	1525	Fl	Fl	Fl	0.689	Fl	0.728	Sh	0.697	?
III.iia	1663	Sh	Sh	Sh	1.000	Sh	1.000	Sh	1.000	Sh
III.iib	2185	Fl	both	Fl	0.999	Sh	0.976	Sh	0.891	?
IV.i	999	Fl	both	Sh	0.996	Sh	1.000	Sh	0.988	Sh
IV.ii	1431	Fl	both	Sh	0.895	Sh	0.867	Sh	0.770	Sh
V.i	1507	Sh	Sh	Sh	0.990	Sh	0.688	Sh	0.931	Sh
V.ii	296	Fl	Fl	Fl	0.991	Fl	0.999	Fl	0.996	-short-
V.iii	1550	Fl	Fl	Sh	0.506	Fl	0.783	Sh	0.952	?
V.iv	807	Fl	Fl	Sh	0.999	Sh	0.997	Sh	0.989	Sh
V.v	653	Fl	Fl	Fl	0.997	Fl	0.615	Fl	1.000	Fl
Pro.	132	Fl	?	Fl	1.000	Sh	0.559	Fl	1.000	-short-
Epi.	268	Fl	?	Fl	0.998	Sh	0.768	Fl	0.997	-short-

*Attribution of each scene according to Spedding [152] and Hoy [55].

Table 7–4: Classification results for *Henry VIII*

each part tested independently. The marker words were counted and program KERCON used to analyze the rates. Table 7–4 provides the results. The format of the table is identical to the earlier table for *TNK*, except that Spedding and Hickson's scene attributions are listed in addition to Hoy's.

Only V.ii, the prologue and the epilogue contain fewer than 500 words. Although the probabilities for these three are listed for the curious, they are too small for the results to be interpreted. The same criteria used in accepting, rejecting and classifying scenes in *TNK* is used for *H8*. Of the 17 samples of at least 500 words, 6 (35%) cannot be assigned to either author. This proportion is higher than was encountered in *TNK* (4 of 17) or in the test set (8 of 88, 9.1%).

Act I

The pattern of occurrence of function words in the first two scenes of Act I strongly resembles Shakespeare's work. The Fletcher z-scores for *in, of, the* and *which* are all above 2.0 in the first scene and greater than 2.5 in the second. The infrequent Shakespeare markers "Infreq-Sh+" are also common: $z_{\mathrm{F1}} = 2.8$ for I.i; $z_{\mathrm{F1}} = 3.2$ for I.ii. These results strongly support the accepted attribution of these scenes to Shakespeare.

The results for I.iii do not agree with the generally accepted view. The posterior probability for set T2 is below 0.80 and rejected, but the the scene is assigned to Shakespeare because the other two probabilities are accepted and the three classifiers agree on the author. The values are not nearly as high as for the first two scenes. While set FR1's value 0.802 is extremely close to being rejected (in which case the scene would have been left unassigned), recall that the reject threshold being used is fairly stringent.

The main factor leading to the assignment to Shakespeare appears to be the 16 occurrences of *of*, which result in a rate of 27.3 ($z_{\mathrm{F1}} = 3.2$). This is extremely untypical of Fletcher; of the 106 scenes in the design set, only 1 has a rate this high, and only 6 more have values greater than 20.0. Fourteen percent of

the Shakespeare scenes have higher rates than this scene. The rate of *in* (5.1) is somewhat low for both authors ($z_{\mathrm{Fl}} = -1.0$, $z_{\mathrm{Sh}} = -1.8$); this is unusual because *in* and *of* are positively correlated in both authors' samples. The probability for set T2 is rejected, and both authors' pooled set of infrequent markers favor Fletcher slightly: for "Infreq-Fl+" $z_{\mathrm{Sh}} = 1.4$; for "Infreq-Sh+" $z_{\mathrm{Sh}} = -1.5$. The result for set FR1 would be more like Shakespeare except for the 7 occurrences of *are* ($z_{\mathrm{Fl}} = 1.5$, $z_{\mathrm{Sh}} = 2.8$), a Fletcher marker unique to this set. These occurrences are scattered throughout the dialogue which forms the bulk of the scene, a discussion of the court gallants and their habits. Examination of the text shows no reason for the relatively large number of occurrences of *of*.

Recall that Mincoff [105] called this scene "the most unmistakably Fletcherian scene in the the whole play" and used it to illustrate a number of Fletcher's characteristics. The small number of verses in this scene have a proportionally large number of feminine endings (which indicate Fletcher) but a high proportion of run-on lines (which point to Shakespeare). It is difficult to evaluate the significance of Mincoff's examples of "typical" Fletcher humor, since Foakes glosses a contemporary parallel of "fool and feather" and a court reference to "fights and fireworks" in connection with the wedding celebrations for Princess Elizabeth. However, the scene does contain several well-known Fletcher constructions: the use of a phrase "and a [a*djective*] one" (line 52) and the use of *else* at the end of a clause (line 65). Table 7–3 shows that other evidence supports Fletcher's claim to this scene; the proportions of *'em* and *ye* are not paralleled in any other Shakespeare scene.

On the other hand, as noted in Section 5.3, the occurrences of *wherewithall* in line 59 and *thereunto* in line 27 are extremely unusual in Fletcher's work. Only 15 of 106 Fletcher scenes in the design set contain any occurrences of a *there-/where-* compound; only one contains 2 occurrences, and the highest rate is in IV.i of *The Island Princess* (1 occurrence, $r = 1.4$). The rate in *H8* I.iii, 3.4 per thousand words, is very unlike Fletcher's normal habit. One possible origin for these words is the source material for the scene. Examination of the extracts

from Holinshed's *Chronicles* reprinted in Foakes' [130] and Schoenbaum's [127] editions do not reveal any such occurrences in the passages behind I.iii (although occurrences of *where-/there-* compounds do occur in other passages).

In conclusion, the overall rates of occurrence of the marker words suggest Shakespeare, although not so strongly as in the first two scenes of Act I. The fact that characteristics of Shakespeare's vocabulary exist side by side with Fletcher's linguistic traits suggests that I.iii represents the work of both men. I believe that the case for revision by one or the other writer is strong. Such a contention would be less easy to defend if the only evidence were the posterior probabilities for the three word sets, but the occurrence of *wherewithall* and *thereunto* provides additional support.

Happily the final scene of Act I provides no shocks. The discriminant analysis procedure assigns the scene to Fletcher, even though the probability for set T1 is just below the reject level. The Shakespeare z-scores for *all, now, are* and "Infreq-Fl+" are all higher than 2.0.

Act II

In Act II one arrives in interesting territory, since Hoy proposes that Shakespeare is responsible for the entire act and that Fletcher merely touched up the first two scenes. The marker words in Scene i do not provide strong indications for either author. The rates for *all* and *dare* favor Fletcher's claim ($z_{\mathrm{Sh}} = 2.3$ and 2.0), but these two do not seem to be strong enough influences to overcome the somewhat Shakespeare-like rates for *the* ($z_{\mathrm{Fl}} = 1.9$) and *of* ($z_{\mathrm{Fl}} = 1.6$). *All* in particular seems important to T1's 0.93 probability for Fletcher. It is not included in set T2, and without it the results for this set favor Shakespeare (although not strongly and in spite of the addition of a non-Shakespearean rate for "Infreq-Fl+," $z_{\mathrm{Sh}} = 2.3$). The results for this scene by no means support Fletcher's claim, nor do they lend any great support to Hoy's case for Shakespeare's responsibility.

The latter may not be true for II.ii. This is the first sample in the two plays that cannot be classified due to two non-rejected probabilities that indicate different authors. In the analysis of scenes of joint authorship in Section 6.5.6, such a result was produced for 5 of the 20 samples, a much higher proportion than for the test set (only 3 of 88). In II.ii, the rates favoring Fletcher are again *all* ($z_{\mathrm{Sh}} = 2.3$) and *dare* (2 occurrences, $z_{\mathrm{Sh}} = 2.0$).[9] The number of infrequent Fletcher markers "Infreq-Fl+" is quite low ($z_{\mathrm{Fl}} = -2.0$), and the rates of *of* ($z_{\mathrm{Fl}} = 2.4$) and *the* ($z_{\mathrm{Fl}} = 1.4$) are partly responsible for the Shakespearean result for sets T2 and FR1. The theory that Fletcher only revised this scene seems to be justified. (Note that the scene contains one of the occurrences of *hath* that puzzled Partridge.)

II.iii also cannot be assigned to either author. Set T1's probability of 0.90 for Shakespeare appears to be heavily influenced by 5 occurrences of *which;* three of these occur in a cluster in lines 28–30. The accepted attribution to Shakespeare is also supported by *of,* but this word does not seem to balance Fletcher rates for *all* ($z_{\mathrm{Sh}} = 1.9$), *too* ($z_{\mathrm{Sh}} = 2.0$) and "Infreq-Fl+" ($z_{\mathrm{Sh}} = 1.5$). If pressed to make an assignment, one might be tempted to go against the accepted view by explaining the T1 result as being mainly due to a rhetorical repetition of *which.* Such a claim would gain more support if one could attribute the low proportion of *ye* to *you* to compositor interference; however, the pages of this scene were set by the workman who appears to have reproduced this Fletcher trait elsewhere. Disappointingly, none of these rates are extremely different from either author's mean rate, and the best decision is to leave the sample unclassified.

[9] *Too* also plays a part ($z_{\mathrm{Sh}} = 1.7$). Two of the five occurrences are found in a repetition in the important exchange:

> *Cham.* It seemes the Marriage with his Brothers Wife
> Ha's crept too neere his Conscience.
> *Suff.* No, his Conscience
> Ha's crept too neere another Lady. II.ii.16–18.

The analysis of Act II closes on a more successful note. The fourth scene, Katherine's trial, has almost always been accepted as Shakespeare's. The function word analysis concurs with this claim. High rates for *in, of* and *which* (for each $z_{\mathrm{Fl}} > 2.3$) make this scene very unlike the Fletcher samples in the design set.

Act III

Such success is short-lived, however, since the examination of Act III begins with another scene almost always attributed to Fletcher that cannot be assigned. The probabilities for all three sets of words are below the rejection threshold. As in IV.i and V.iii in *TNK*, few of the word rates in this scene are very different from either author's expected rate. For sets T1 and T2, all but two of the z-scores for both authors have absolute values less than 1.2. The most positive evidence for Fletcher's authorship is 3 occurrences of *dare,* which produce a z-score for Shakespeare of 2.0.

The first part of III.ii was accepted as Shakespeare's by Spedding and by all scholars since. Program KERCON also assigns this sample to Shakespeare. The strength of the word-rate evidence in this sample is as strong as the results for *TNK* I.i–iii, III.i, V.i and *H8* I.i–ii. The Fletcher z-scores for *in, the* and *which* are all above 3.0, while *of* also contributes with a value of 2.0.

The second part of III.ii, which is usually assigned to Fletcher, is another scene that Hoy believes was only revised by the younger man. The posterior probabilities for this scene are most striking; all three are high (two are greater than 0.97) but sets T1 and FR1 classify the sample as Shakespeare's while set T1 assigns it to Fletcher. Set T1's result can be attributed to a large number of occurrences of *dare.* In fact, 7 of the 20 occurrences in the entire play can be found in this sample (which contains 9.2% of the total number of words in the text). The resulting rate of 3.2 is very different from the Shakespeare average rate ($z_{\mathrm{Sh}} = 5.3$). Set T1's high probability for Fletcher is also helped by a high rate for *all* ($z_{\mathrm{Sh}} = 2.8$). This word is also part of set FR1, but it does

not affect the assignment there. Rates for *of* ($z_{\mathrm{Fl}} = 2.7$) and *the* ($z_{\mathrm{Fl}} = 2.2$) appear Shakespearean enough to ensure that FR1 assigns the sample to the elder playwright. The extremely different results for sets T1 and T2, which both include *dare,* are clearly due to the fact that the rates for the infrequent pooled marker sets are very unlike Fletcher. Both of these variables favor Shakespeare; $z_{\mathrm{Fl}} = -1.4$ for "Infreq-Fl+", and $z_{\mathrm{Sh}} = 2.1$ for "Infreq-Sh+". Together with the frequency of *of* and *the,* these rates result in a probability for T2 that supports Shakespeare's authorship of III.iib as decisively as T1's result supports Fletcher's.

In this scene, Wolsey realizes that he has fallen from the king's favor. He is confronted by his enemies, who demand that he surrender his seal of office. He refuses, and after a heated confrontation he laments his fall. Of the 7 occurrences of *dare*, 5 occur during the exchanges between the cardinal and his enemies: for example, "Who dare cross 'em [the king's orders]" (line 234); "I dare and must deny it" (line 238); and "I dare your worst objections" (line 307). The high rate is in part due to the subject matter, but I think that there are still too many occurrences of *dare* to attribute them all to a Shakespearean stylistic effect.

Of course one cannot "remove" occurrences of a word, but this provides a good opportunity to play "what-if" with the kernel classifier. If there were 2 fewer occurrences of *dare* in III.iib, then $z_{\mathrm{Sh}} = 3.6$ and the posterior probability for Fletcher is still 0.97. If there were 3 fewer, then $z_{\mathrm{Sh}} = 2.8$ and the resulting probability 0.73 still favors Fletcher but is below the rejection threshold. If there were only 3 occurrences rather than the actual 7, then $z_{\mathrm{Sh}} = 1.2$ and the probability for set T1 now begins to indicate Shakespearean authorship at 0.60. Thus one or two occurrences of *dare* do not change the T1 result. These results indicate that large samples like III.iib (which contains 2185 words) may be relatively immune to the addition of a few occurrences of a strong but infrequent marker like *dare.*

In summary, an unusual proportion of *dare* suggests Fletcher's presence, but this evidence is not so strong that the result is maintained when "Infreq-Fl+" and "Infreq-Sh+" are included. The result for set FR1, which does not include any

of these three, also indicates that the sample is different from Fletcher's unaided work. I believe that these results support Hoy's contention that Fletcher revised Shakespeare's work in the second part of III.ii.

Act IV

Act IV of the play is also controversial. Again, Hoy believes that the clusters of *ye* in the two scenes are due to revision by Fletcher, although Schoenbaum rejects this conclusion for the first scene [127, p. xxxvi]. Even Spedding's first reaction to Act IV was that he "did not so well know what to think;" the speeches seemed too vigorous for Fletcher but lacking the freshness and originality of Shakespeare [152, pp. 9*–10*]. It appears that Hickson's opinion and the metrical test results convinced him of Fletcher's authorship.

Discriminant analysis of the three sets of marker words assigns both scenes to Shakespeare. The high probabilities for IV.i are due to extremely large rates for *the* ($r = 69.1$, $z_{\mathrm{Fl}} = 6.3$, $z_{\mathrm{Sh}} = 3.8$) and *of* ($r = 35.0$, $z_{\mathrm{Fl}} = 4.9$, $z_{\mathrm{Sh}} = 2.2$). The combination of such high rates is unlike anything in the design set; only 3 Shakespeare scenes out of 265 have a rate for *of* greater than 30.0 in conjunction with a rate for *the* greater than 50.0. These rates are partially due to eleven occurrences of *the* and 10 or 11 occurrences of *of* that are part of the titles of the participants in the coronation procession (for example, "the Duke of Suffolke"). Thus these unusually high rates are partly a product of the subject matter, but even if one completely ignored these occurrences, the rates for these two words would still be very unlike Fletcher.[10] The rate for *which* also supports Shakespeare ($z_{\mathrm{Fl}} = 2.8$), while rates for *are* and *all* indicate Fletcher ($z_{\mathrm{Sh}} = 1.7$ and $z_{\mathrm{Sh}} = 2.9$). The multivariate combination is on the whole much more like

[10]The scene contains 999 words, so the rates for *of* and *the* would be 58 and 24 if the titles were completely ignored. Fletcher's mean rate for *of* is 12.6 with a standard deviation of 4.6; for *the,* his mean is 23.4 with a standard deviation of 7.2. The z-score for *of* would still be greater than 2.5, while the value for *the* would be greater than 4.

Shakespeare. If Fletcher made minor contributions, they may have involved a few of his markers (like *all* and *are*).

IV.ii is assigned to Shakespeare by all three word sets. The probability for set FR1 is 0.77 and rejected, and the other two probabilities are not as high as for I.i–ii and III.iia. In contrast to the preceding scene, the rates for *the* and *of* are not very different from either author's average rate. Rates for *which* ($z_{\mathrm{Fl}} = 2.2$), *in* ($z_{\mathrm{Fl}} = 1.7$) and "Infreq-Sh+" ($z_{\mathrm{Fl}} = 2.2$) point towards the writer from Stratford, while *now* ($z_{\mathrm{Sh}} = 1.5$) is the only word that occurs at a rate more like Fletcher's mean.

The analysis of the marker words in Act IV indicates that Shakespeare was responsible for both scenes. Thus Katherine's final speeches cannot be fully credited to Shakespeare's collaborator. If this scene raises difficulties in the interpretation of the play's structure, as some critics maintain, then their existence cannot be explained away as Fletcher's misunderstanding of Shakespeare's intentions.

Act V

The final act begins with a scene generally agreed to be Shakespeare's. My results agree with this view, although the probability for set T2 is rejected. This rejection is somewhat surprising when the z-scores for that set are examined. For "Infreq-Sh+," $|z| < 1.0$ for both writers, and the rate of "Infreq-Fl+" does not appear to weaken Shakespeare's case ($z_{\mathrm{Fl}} = -1.6$, $z_{\mathrm{Sh}} = -0.3$). The scene's 2 occurrences of *dare* ($z_{\mathrm{Sh}} = 1.9$) seem to be the cause (although this word is also included in T1). These occur in lines 38–39 in a repetition linked with the subject matter: "who dare speak One syllable against him? ... There are that Dare...." Despite this Fletcher-like rate for *dare*, the occurrences of *of* ($z_{\mathrm{Fl}} = 2.2$), *which* ($z_{\mathrm{Fl}} = 2.5$) and *the* ($z_{\mathrm{Fl}} = 1.5$) apparently ensure that each classifier allocates the scene to Shakespeare.

The three probabilities for V.iii cannot be used to assign the scene depicting Cranmer's trial and rescue (usually given to Fletcher) to either author. Set FR1 indicates Shakespeare with a probability of 0.95; the rates in this set appear to be dominated by a very high value for *of*. The 39 occurrences ($r = 25.2$, $z_{\mathrm{Fl}} = 2.7$) include 8 titles ("his grace of Canterbury," "my lord of Winchester"), which might help explain the non-Fletcher rate. Another marker with a Shakespearean rate in the scene, *which,* is repeated in reference to Cranmer's purported heresies in lines 18 and 20. On the other hand, 2 of the 3 occurrences of *dare* occur when Cranmer's enemies discuss the men "who dare accuse you" in lines 50 and 56. The rates for the more common markers point in both directions; *are* occurs more often than normal in Shakespeare ($z_{\mathrm{Sh}} = 2.7$, $z_{\mathrm{Fl}} = 1.4$) and *in* is somewhat unlike Fletcher ($z_{\mathrm{Fl}} = 1.5$, $z_{\mathrm{Sh}} = 0.0$). For both sets of infrequent markers, $|z| < 0.9$. Examination of the z-scores is once again unhelpful; it simply confirms what can be inferred from the three probabilities: the rates in V.iii only provide weak and conflicting evidence.

The results for V.iv, on the other hand, are much stronger; in addition, the marker-word evidence contradicts the accepted assignment to Fletcher. The probability of each of the three classifiers is greater than 0.98. In this scene, the porter and his man confront the mob that threatens to force its way into the court to witness the christening of the baby Elizabeth. As noted on page 316, Mincoff feels that the scene is "very unmistakably" Fletcher's despite the metrical evidence and the fact that it is mainly prose. The high proportions of *'em* and *ye* are sounder evidence. The proportions for this scene, indicated in Table 7–3, are unparalleled in Shakespeare's known work.

The scene contains several very unusual rates for the marker words used in this study. The rate for *the* is 49.6 ($z_{\mathrm{Fl}} = 3.6$, $z_{\mathrm{Sh}} = 1.8$). Only 2 of the 137 Fletcher scenes in the design and test sets have rates above 45, in comparison to 34 of the 322 Shakespeare scenes. But several words have marked Fletcher-like rates. *Are* occurs very frequently ($r = 14.9$, $z_{\mathrm{Fl}} = 2.4$, $z_{\mathrm{Sh}} = 3.9$), and a number

of occurrences of *'em* and *them*[11] help produce a rate for "Infreq-Fl+" that is very unlike Shakespeare ($z_{\mathrm{Sh}} = 3.0$). But the rate for "Infreq-Sh+" is somewhat higher than the Fletcher average rate ($z_{\mathrm{Fl}} = 1.6$). The subject matter in this scene is probably affecting the rates of *are* and "Infreq-Fl+." The two mens' discussion of the crowd will naturally involve an increased use of the third-person plural forms of pronouns and *to be*.

Although a number of rates make V.iv an unusual scene for both playwrights, it is far more unlike Fletcher than Shakespeare. One might postulate that Fletcher's use of function words changes very dramatically in his prose, but since he rarely writes prose, this might be difficult to prove. I think one must accept the conclusion indicated by the discriminant analysis results (within the limits of their demonstrated accuracy on the design and test sets), and accept the procedure's assignment of the scene to Shakespeare.

The final scene in the *H8* contains Cranmer's prophecy praising Queen Elizabeth. The procedures used in this study assign the scene to Fletcher, which agrees with accepted opinion. The frequency of "Infreq-Fl+" ($z_{\mathrm{Sh}} = 3.0$), of *must* ($z_{\mathrm{Sh}} = 3.2$) and primarily of *all* (11 occurrences, $r = 16.8, z_{\mathrm{Fl}} = 2.2, z_{\mathrm{Sh}} = 5.4$) are significantly unlike Shakespeare's normal use. Three of the five occurrences of *must* are found together, in what many will regard as a typically parallel Fletcher construction: "But she must dye, She must, the Saints must haue her" (lines 60–61). Repetitions of this word within a few lines are very prominent in *TNK* II.ii and *TNK* IV.ii.

7.2.3 Summary

A reader with a firm belief in the accepted divisions of these two plays will probably be less satisfied with the results for *H8* than for *TNK*. Almost all the scenes

---

[11] Recall from Section 5.5.1 on page 216 that this pair, when both forms are counted together, is a useful Fletcher marker.

generally accepted as Shakespeare's (I.i–ii, II.iv, III.iia and V.i) are marked by function word rates that are very unlike scenes by Fletcher. The only exception is II.iii, which cannot be assigned. These results also give some support to Hoy's conclusions that Fletcher only revised some other scenes that are basically Shakespeare's. Both scenes of Act IV resemble Shakespeare's design-set samples. In addition, the "high-but-disagreeing" posterior probabilities for II.ii and III.iib are similar to the pattern found in one-quarter of the samples of joint authorship tested in the last chapter, suggesting that these scenes could well be the work of both authors. (One should also not forget that the same study showed that 50% of the joint samples were assigned to one author or the other.)

The positive indications of Shakespeare's hand in I.iii and V.iv are more controversial. The probabilities of authorship for I.iii are not as strong as those of several scenes accepted as Shakespeare's. But this result is also supported by other lexical evidence: the occurrences of *wherewithall* and *thereunto*. However, the presence of Fletcher's stylistic traits is more convincing in I.iii than in many other scenes, and revision seems the most easily accepted conclusion. The use of marker words in V.iv is unusual for both writers, but much more so for Fletcher. The statistical results are quite strong, but accepting that Shakespeare wrote this scene raises an extremely awkward and important question: Why does the Folio text contain high proportions of *ye* and *'em* in a scene by Shakespeare? Since scholars have shown conclusively that these orthographical features are subject to alteration by scribes and compositors, Foakes' assertion that too little is known about the copy for the Folio text is perhaps not "merely frivolous," as Mincoff states.

The kernel classification results for *H8* resemble those for *TNK* in that no scenes are attributed to Fletcher with high probabilities. In fact, the rule used to derive a verdict from the three probabilities only gives him two scenes, I.iv and V.v. Two scenes which are often considered to contain the strongest stylistic evidence for his presence in *H8*, III.i and V.iii, are not assigned to either author by this procedure.

The inability to recognize scenes by Fletcher with the same certainty as those by Shakespeare may be a feature of the classifiers based on these three word sets and the kernel method. It could well tie in with the fact that there are fewer Fletcher scenes in the design set, or that the three most frequent words used (*the, of* and *in*) are Shakespeare markers. When classifying the test set, the posterior probabilities calculated by program KERCON are greater than 0.90 for all three sets of words for 37 of the 57 (65%) Shakespeare scenes but only 12 of the 31 (39%) Fletcher scenes. Of course, these proportions may simply reflect the characteristics of the six test-set plays, but there is a suggestion that the entire procedure is better at recognizing Shakespeare than Fletcher.

While this may indicate that a number of the rejected scenes attributed to Fletcher really are his, it should not raise doubts about results that indicate Shakespearean authorship. The likelihood of an error in assignment is best evaluated by misclassification results for the design and test set discussed in Section 6.5.5 on page 286. In addition, z-scores have been examined in detail in an attempt to determine how individual words affect a particular classification result. While this has often been useful, it should be re-emphasized that judgements regarding authorship should not be made from these univariate statistics. The kernel classification method considers any relationships between words when weighting each individual variable's contribution to a result. It also bases its result on the number and proximity in the measurement space of design-set scenes with similar rates.

Chapter 8

Discussion

This concluding chapter consists of a review and discussion of some of the findings of this study. These are examined in relation to the techniques and features of several other authorship studies. Some possible criticisms are anticipated, and possible avenues for future research are mentioned. In this review, I will address issues involving the texts used in the study, the variables examined, the statistical methods evaluated and, finally, some of the implications for authorship study in the area of English Renaissance drama.

8.1 Textual Considerations

One strength of the current study is the amount of text examined. The 20 plays of the Shakespeare design set contain 421,622 total words (before the expansion of compound contractions); the 4 plays in the test set provide an additional 89,145 words. Fewer Fletcher texts were available, but the 6 plays in the design set together total 130,879 words; the two test set plays contain a total of 89,145 words. While sampling methods could have been used to avoid dealing with large volumes of data, I believe that it is better to err on the side of safety. The three plays examined by Michaelson, Morton and Hamilton-Smith in "To Couple Is the Custom" gave no hint of the sizeable within-author variation of collocations that

is observed in 20 Shakespeare plays. The establishment of habits of authorship should be based on as many samples as possible. Studies that use only two or three samples to represent an author (for example, O'Brien and Darnell's *Authorship Puzzles in the History of Economics*) are not particularly convincing.

The amount of text used compares favorably to other studies. As noted at the end of Chapter 4, the authors of "To Couple Is the Custom" established their methods on about 215,000 words of known authorship (before applying them to several authorship problems, for which they made use of smaller amounts of undisputed text). Samples from two novels by Scott formed the largest sample by a single author, about 115,000 total words. Austin, in his examination of *The Groats-worth of Wit* [3], used about 100,000 words of Chettle and 40,000 words of Munday for his control samples. Mosteller and Wallace's analysis of *The Federalist* papers was based on papers known to be by Hamilton totaling 94,000 words and papers by Madison totaling 114,000 words. (In addition, they used a number of other samples to select markers and to study variation with time in both writers.)

Of course one feels less secure about the integrity of Shakespeare's and Fletcher's plays than about *The Federalist* papers or Scott's novels. Sceptics (troublesome supervisors, for example) can quite justifiably ask, "Are you certain that all the control texts are *really* by Shakespeare or by Fletcher?" To begin an authorship study one must make some initial assumptions. Each result in this study depends on the validity of my assumption that the plays listed in Table 2–1 (page 39) are by a single man (who, for convenience's sake at least, we can call Shakespeare), and that the 6 plays selected from those in Table 2–2 (page 45) are by Fletcher. I think these assumptions are reasonable, but others may have different views. A large majority of scholars would agree with my opinion. Almost all would point out that small parts of any of these plays might be corrupt or by another hand. As noted in Chapter 2, any text recognized to have major textual problems was not used. The best texts that have survived (as far as we know) have been chosen. If one has doubts about these, then it is

impossible even to question the integrity of *Henry VIII* or to consider how best to test the possibility of collaboration in either play.

The choice of texts is of fundamental importance, and a poor decision can undermine the most thorough or sophisticated analysis (for example, Austin's study of *The Groats-worth of Wit*, discussed at the end of Chapter 3). People reading about the current study in a few years may well remark, "He should have used the Oxford critical old-spelling edition of Shakespeare." It is almost annoying that this important edition has been published just as I finish my research. However, the Shakespeare texts I have chosen are relatively "clean" texts, and I believe that most (if not all) of the counts presented in this study would not differ much from those made from the new critical edition. On the other hand, if this edition had been available to me in machine-readable form, I could have considered using some of the plays that were left out of the study because of textual problems (for example, *2 Henry IV* and possibly *Othello* and *Lear*.) Shakespeare's tragedies have not been completely neglected in this research, but the Oxford edition could be used in further research to test some of the great Shakespearean tragedies which were not examined.

In addition, another volume of Bowers' *The Dramatic Works in the Beaumont and Fletcher Canon* has been published since I began this study. This volume includes at least 4 plays that are suitable for use as Fletcher controls. Obviously it would desirable to use as many texts as possible in an application of discriminant analysis. If these additional plays were available, one could use the analysis of variance procedure to examine the internal variation within Fletcher's texts more closely (as was done for Shakespeare in Chapter 5). Also, it would be very interesting to use the procedures developed in Chapter 6 to examine *The Faithful Shepherdess*, mentioned in a footnote in Section 2.4.1. This pastoral work is written in an archaic style most unlike Fletcher's other work, and it would be illuminating to see if such a change in style has greatly affected the rates of function words.

8.2 Choice of Variables in an Authorship Study

8.2.1 Positional Stylometry versus Frequency Alone

This study began four years ago as an application of positional stylometry. After the findings presented in Chapter 4, the position of words plays no further role in the examination of the authorship question. The only type of "positional" variable evaluated in this study is the collocation. (Although proportional pairs were discussed in the context of positional stylometry, they are based on the ratios of word frequencies and really are not positional variables at all.) The basic conclusion of Chapter 4 is that there is nothing magical about collocations. Those tested do not have an unusually stable rate of occurrence in samples of Shakespeare and Fletcher. The *t*-tests show that the two playwrights differ more in how often they use a number of common words (once within-author variation is taken into account). This result for collocations and their component words has only been demonstrated in Shakespeare and Fletcher. Other studies using collocations (including "To Couple Is the Custom" and *Literary Detection*) do not include comparisons of collocations with the rates of common words. It would be interesting to assess the usefulness of word rates in some of the problems addressed in these works.

The basic premise of positional stylometry (that variables based on the combination of frequency and position are better indicators of authorship than variables based on frequency alone) is very attractive. The studies of Greek texts published by Morton, Michaelson and their associates demonstrate the great number of possible ways of measuring frequency and position. Some of these measures could be applied to 16th and 17th century dramatic texts, and other definitions of position (such as the relative position in verses or speeches) could be tested. New positional variables were not investigated in this study, not because I reject

the basic premise of the approach of positional stylometry, but because the study of word frequencies appeared to have potential.

The study of grammatical word class in Jacobean texts may lead to valuable tests of authorship. As noted several times in this dissertation, accurate and efficient software that automatically recognizes a word's part of speech has been developed for 20th century English. Such a study of Jacobean plays must wait until such software is adapted to Early Modern English, since it would be a major undertaking to tag the large number of texts required to establish an author's pattern of usage by hand. Both the positional and frequential approaches to stylometry will certainly be applied to word-class variables in the future. The high frequency of many of the classes may result in useful variables for authorship study.

8.2.2 Using Word Rates as Variables

One of the most remarkable results of this study is that small samples of Fletcher and Shakespeare can be distinguished using words like *all, in, of, the* and *too.* When Austin includes an examination of frequent function words in his study of *The Groats-worth of Wit,* he remarks: "It was not to be expected that any two writers would vary greatly in their use of this linguistic small change" [3, p. 25]. The results of my study show that this is as untrue for two Jacobean dramatists as it is for Hamilton and Madison. Some of the "filler" words of English are extremely useful in resolving a difficult and controversial Shakespearean authorship question. One should not forget that these function words were not randomly or subjectively chosen, but by using statistical tests to evaluate their potential for discrimination. Samples of other Jacobean writers should be examined to determine if these same function words are often used at such different rates.

A major advantage of using function words to study authorship is their high frequency. The importance of using frequent markers was discussed in the Introduction and in Section 3.1.2. Function words occur more often than any of the collocations examined in this study. One result of the statistical advantages

of frequent markers is that smaller samples can be examined. This has been particularly important in this study, since one goal was the independent evaluation of internal evidence in individual scenes of *Henry VIII* and *The Two Noble Kinsmen*. The total rate of occurrence of the variables used in classification in the control texts is fairly high for both authors. Twelve of the sixteen markers were used in at least one of the three subsets of words, T1, T2 and FR1. In the 20 Shakespeare plays the combined rate for these twelve is 105.5 per thousand; in the 6 Fletcher texts the rate is 99.8. Thus, about 10% of the total number of word occurrences in the texts are used in the classification procedure.

The examination of common words in Chapter 5 had the sole goal of identifying words and classes that were useful markers of authorship. However, several of the analyses in that chapter could be expanded into complete studies in their own right. The ANOVA tests revealed some interesting differences in the use of function words between groups of Shakespeare's texts. In addition, correlation between pairs of words could be explored further in Jacobean and modern literary texts. An analysis of word use and characterization is another area where computers and statistics could be applied.

8.2.3 Identifying Occurrences of Words

A great deal of effort has gone into producing versions of texts in which all occurrences of common words can be identified and counted. While homonyms occur in most texts, the proportion of compound contractions in English Renaissance drama is higher than in texts from most other periods and genres. That compound contractions should be expanded to their full forms (before counting function words) can be justified on the basis of the possibility of non-authorial revision by scribes, printers, editors or revisers. But I think expansion is desirable on linguistic grounds. If one is counting occurrences of *it* or the verb *to be,* then why should occurrences of *it's* and *'tis* be ignored? If one wishes to recognize an author's preference for contracted or full forms, it makes more sense to study contraction rates directly (as was done in Section 2.6).

Again, expansion will not affect word counts very much for most types of literary text. But the analysis in Section 2.6 indicates that contraction must be examined in studies of Jacobean drama. Before comparing how two authors use common words, one must decide how to deal with differing rates of contraction between authors and secular changes within the works of a single author (as occur in Shakespeare). Although I believe that my decision to expand compound contractions is justified, expansion is a large obstacle for any other researcher who wishes to validate my findings or apply my methods to other problems. I would be very pleased if the translation lists and replacement software I have developed could be put to use by others in such studies.

8.2.4 Contextuality

Concern regarding whether marker words are context-free or not been a major concern in other authorship studies. (Mosteller and Wallace's discussion of the contextuality in their summary of conclusions [113, pp. 265–266] outlines the problem.) I have paid less attention to this problem than others, which many may regard as a weakness of this examination. I have hoped that any serious problems in this regard would manifest themselves in the analyses of the samples in the design and test sets (especially in the feature selection methods used in Section 6.4). It is not clear whether or not this approach has been successful, especially in the case of *dare* and *Henry VIII* III.iib. In choosing variables a researcher is faced with a dilemma: should one select a word that is usually an excellent marker but is occasionally affected by style or subject matter? *Dare* can occur as an auxiliary verb, a lexical verb or a noun, and one might imagine that the last two forms might be less context-free than occurrences of the auxiliary in Early Modern English.

But how can one determine the extent of context-dependence? There is no good answer. Austin and Mosteller and Wallace eliminated words that they feared might be affected, but examination of the words they have chosen shows that these subjective decisions admit some doubtful (to my mind) markers (such

as *city, language, aim* and *admire*). An objective method of evaluating the extent of contextuality is required, but I find it impossible to accept Damerau's definition that significant contextuality is determined by a non-Poisson pattern of occurrence (discused in Chapter 3). Other means of objectively assessing dependence on context should be examined.

Discriminant analysis may provide one way of recognizing samples of known authorship with extremely unusual word rates. Silverman discusses the use of an *atypicality index* as a measure of how representative of its class an observation is [142, p. 128]. This measure is based on the probability that a randomly chosen observation from the class will have a pdf value greater than that of the suspected outlier. Apparently the potential of atypicality indices has not been fully realized in statistical studies, possibly because their calculation usually requires multivariate integration over the measurement space. However, they could be very useful for recognizing samples that were extremely unlike other samples by the same author. This would be especially useful for assessing the evidence in scenes like *H8* III.iib, where the subject matter may be affecting rates, and *H8* V.iv, where the observed rates are unusual for both candidates.

8.3 Statistical Methods

8.3.1 Measuring the Discriminating Power of Individual Variables

The first step in an authorship study is to choose the features that best discriminate between the candidates. When choosing from a large set of words, it is important to find a statistical measure for recognizing potential markers quickly. Two measures were used in Chapter 5 to assess the value of an individual word: the distinctiveness ratio and the t-test. The distinctiveness ratio, used by Ellegård and Austin, has two disadvantages. First, it takes no account of

within-author variation. Second, a word with a moderate or high rate in both authors may not have a large distinctiveness ratio, although the difference between the rates is significant. But this value was useful in this study for recognizing low-frequency markers (or words that one author very rarely uses).

The t-test seems a natural choice as a measure of discriminating power, since it measures the difference in mean rates in terms of the within-author variation. (Note that the use of the t-test in this study has a slightly different purpose than in most situations, where it is used simply to determine if two means are significantly different.) The t-test is more difficult to calculate, since measurements on a number of samples are required to determine an estimate of the variance for each author. Some might see this as a disadvantage, but it does force one to consider within-author variation, a factor that has been ignored too often in authorship studies.

8.3.2 The Use of χ^2 Tests

Many of the positional stylometry studies reviewed in Chapter 3 used a testing procedure based on χ^2 tests to make judgements regarding authorship. Most often an $n \times 2$ contingency table was used to compare counts of features in a disputed sample (say, an act) to the total of the counts made in a set of control samples (say, three plays). (The comments that follow also apply to the use of Fisher's exact test with 2×2 tables.) The Null Hypothesis tested is that the proportion of occurrences in the disputed sample is the same as the combined counts in the control set. Such a testing procedure does not consider the expected variation within classified samples of the same length as the disputed sample. Possibly some researchers have not worried too much about this because of their belief that the features they have chosen do not vary significantly within authors. One only needs to look at the graphs of the within-author variation for collocations and proportional pairs (Table 4–6 on page 160) or function words (Table 5–12 on page 215) to see that such an assumption is probably rarely justified.

Some have defended the use of such an approach on the grounds that the χ^2 test is distribution-free (for example, O'Brien and Darnell [117, p. 16f]). The χ^2 goodness-of-fit test is certainly distribution free, but this does not justify the use of tests on contingency tables as described in the preceding paragraph. In that approach, the χ^2 test (or Fisher's exact test) is used to provide an answer to this question: "How different is the proportion of the observed counts from the best estimate of the expected value, derived from the total counts of a number of samples?" A better question is: "What is the likelihood of observing a sample with these counts, given the pattern of occurrence observed in a number of similar samples?" To study pattern of occurrence, one has to examine the statistical distribution of the features being considered.

There is nothing inherently wrong with using the χ^2 test (or Fisher's exact test) to compare samples. I suggest that there is a flaw in the manner in which many of the studies reviewed in Chapter 3 have made use of it. I think it desirable to compare like to like, and one might be able to make use of χ^2 or Fisher's test to determine the limits of variation for small samples of known authorship. For example, if a disputed text was long enough that a number of samples could be taken from it, the χ^2 goodness-of-fit test could be used to see how well these samples match the frequency distribution of similar length samples observed in an author's known works.

8.3.3 Distribution-free Discriminant Analysis

The comparison of like to like is at the center of discriminant analysis techniques. The multivariate nature of these techniques also eliminates the problem of combining the results of a number of possibly non-independent significance tests. There are some disadvantages to using this approach to study words, however. One is the necessity of measuring word rates rather than counts, which was discussed earlier. Another important problem is the "curse of dimensionality," which stems from representing observations as vectors in a measurement space and then trying to estimate a probability density function from a fixed number

of observations. This puts a fairly small limit on the number of variables that can be used in an analysis.

The distribution-free method used to classify scenes from *Henry VIII* and *The Two Noble Kinsmen*, the fixed kernel estimator, successfully handles many of the distributional difficulties that often characterize textual features. Mosteller and Wallace's approach uses univariate parametric classifiers, and estimating the parameters of the distributions is a very difficult problem (even for one-dimensional pdfs). The counterpart to this problem in the kernel method is the estimation of smoothing parameters for each class. Statisticians have developed several solutions to this problem, and these are often implemented in available computer software (like program KERCON or the ALLOC80 package, described by Silverman as being in widespread use [142, p. 129]). The kernel method appears to be more useful in the analysis of textual data than the other non-parametric method tested, k-nearest neighbor classification, because it allows for a unequal within-class variances for each feature and more easily handles classes with different numbers of observations.

In the preface to *Inference and Disputed Authorship* [113, p. viii], Mosteller and Wallace report some comments made by Neyman in response to the presentation of their research at a conference:

> Neyman suggested that categorizing statistical methods as Bayesian or non-Bayesian is less revealing than categorizing them as inferential or behavioristic, in either of which Bayes' theorem may often be used. The behavioristic approach in our problem calls for establishing a rule for deciding who wrote any disputed paper and evaluating or bounding the frequencies of incorrect classifications if the rule is followed. In the inferential approach, one tries to provide odds or other measures of confidence for (or against) Madison's authorship of any paper.

This points out a major difference between Mosteller and Wallace's main study and the procedures I develop in Chapter 6. By evaluating misclassification and

rejection rates in the design and test sets, this study follows the behavioristic approach.[1]

Future research in the use of distribution-free discriminant analysis in a literary context should examine variable kernel methods. This modification to the kernel approach is designed to produce more accurate pdf estimates for the distribution tails. Silverman describes one such method in detail, the *adaptive kernel,* and notes that the ALLOC80 package includes both fixed and variable kernel procedures. In addition, it is important to determine how well the method performs when fewer classified samples are available, since for many Jacobean authorship questions, fewer undisputed texts by the candidates have survived than for the two authors considered in this study.

8.4 Elizabethan and Jacobean Authorship Questions

Finally, the results of the analysis of *The Two Noble Kinsmen* and *Henry VIII* will be discussed. The evidence presented by function words will be compared to other internal evidence, and the textual issue examined once again. Other studies will be suggested, including further analyses of the nature of collaboration and different authorship questions that might be suited to this approach.

First, it is useful to re-examine the basic approach of this study. The early stages of the analysis of function words were an attempt to answer these questions: Is there a difference in the rate of occurrence of function words in these plays that corresponds to a difference in authorship? If so, how close is this correspondence, and how often does it lead to apparently incorrect decisions about undisputed

[1]Neyman also suggested that the non-parametric discrimination method developed by Fix and Hodges in 1951 (the nearest neighbor method) could be applied to authorship problems. Perhaps it is surprising that this suggestion has only recently been pursued.

samples? These were the issues addressed in Chapters 5 and 6. Equipped with the answers to these questions, the next step is the application of the tests to the scenes in the two disputed works.

But how should one interpret the results? When the classifiers are applied to the word rates from a disputed scene, the "verdict" should perhaps be formulated along these lines:

> The rates for these words in this scene are more similar to the rates found in undisputed scenes written by Author A than those written by Author B.

The only grounds one has for making the jump to the statement "Author A wrote this scene" are the results from Section 6.5.5, which show that such a conclusion was correct for 96.5% of the 365 scenes in the design set and for 87.5% of the scenes in the test set (94.8% overall). The probabilities produced by the discriminant analysis procedures also provide some indication of the certainty of the decision (and a means of recognizing scenes that the method should leave unassigned). This is important, for as Hoy notes, "With linguistic evidence it is all, finally, a matter of more or less" [55, p. 87]. In the scene-by-scene analysis of *TNK* and *H8*, it was valuable to recognize that the evidence for some scenes was much stronger than for others.

8.4.1 The Collaboration of Shakespeare and Fletcher

The results of the analysis of scenes from *TNK* and *H8* are summarized in Table 8–1. This table lists the scenes in both plays according to whether the marker word rates are "very like" one of the author's undisputed scenes in the design set or simply "like" author's scenes. Scenes that were were too short to analyze or that were unassigned by the classification procedures are also listed. The scenes in the "very like" category all have posterior probabilities for all three sets of marker words that are about 95–99%.

Do I believe that the results in Table 8–1 represent the division of authorship in the two plays? Yes, within the limits of error discussed above. I would also

The Two Noble Kinsmen

Very like Shakes.	I.i–iii, II.i, III.i, IV.iii, V.i,iv
Like Shakes.	II.iii, V.iii
Very like Fletch.	II.v, III.iii
Like Fletch.	II.ii, III.vi
Unassigned	III.v, IV.i–ii, V.ii
Too short	I.iv–v, II.iv,vi, III.ii,iv, Pro., Epi.

Henry VIII

Very like Shakes.	I.i–ii, II.iv, III.iia, IV.i, V.iv
Like Shakes.	I.iii, IV.ii, V.i
Very like Fletch.	
Like Fletch.	I.iv, V.v
Unassigned	II.i–iii, III.i,iib, V.iii
Too short	V.ii, Pro., Epi.

Table 8–1: A summary of the classification results for *TNK* and *H8*

qualify this belief by noting that the study of samples of joint composition in Section 6.5.6. Of the 20 samples tested, 6 would fall into the "very like" category of Table 8–1 and another 4 into the "like" group.

I have no personal reasons for wanting to push Shakespeare's claim to *TNK* II.iii. No other critics have recognized his hand in this scene, as far as I know. Perhaps he wrote it, or perhaps this is one of the errors that we expect the classification procedures to produce. (Or perhaps it was written by a third author, although such a possibility is beyond the scope of this study.) I do feel that the evidence for Shakespeare's authorship of IV.iii is strong. In *Henry VIII,* I think that there is evidence of Shakespeare's hand in I.iii, but this must be reconciled with stylistic traits that closely resemble Fletcher's. The marker word rates in Act IV also resemble Shakespeare's scenes rather than Fletcher's, and if the younger writer is present, he must have introduced only minor revisions (a judgement that agrees with the views of Hoy and Foakes). Although II.ii and III.iib are not assigned by the classification rule used, similar results occurred for 25% of the joint composition samples compared to only 3.4% of the test-set

scenes. This, in my opinion, supports Hoy's contention that Fletcher was not entirely responsible for the composition of these scenes.

8.4.2 Comparison to Other Internal Evidence

This study has shown that function words can be studied as internal evidence of authorship, and one should consider how this evidence compares to other forms of internal evidence. This becomes especially important when the results of an analysis of function words do not agree with results based on more traditional forms of evidence. One reason that function words might be a reliable form of internal evidence is that, because of their frequency and lack of prominence, they are presumably less likely to be altered by scribes, printers, editors or revisers. For the procedures used in this study, this will only be true if alterations in the counts (due to corruption or revision) do not affect the word rates enough to drastically change the posterior probabilities. Because of the relation between counts and rates, one or two insertions or deletions of an infrequent marker (like *dare*) can produce a large change in the rate, especially in short scenes. Classification results that appear to be strongly affected by a rate for one infrequent marker are thus less trustworthy than a result due to one or more frequent markers (like *all, the, of* and *in*).

My result for *Henry VIII* V.iv probably represents the most serious disagreement between my analysis of function words and other examinations of linguistic evidence. The marker word rates in this scene are much more like those in Shakespeare's design-set samples than in Fletcher's, but it contains 13 occurrences of *'em* to none of *them,* and 8 occurrences of *ye* to 15 of *you.* If one accepts that my result indicates Shakespearean authorship, one must then explain these very Fletcher-like proportions for the pronoun forms. Either Shakespeare was capable of breaking from his normal practice, or these forms were introduced into the copy by a scribe or Fletcher the reviser. Both of these explanations require a serious departure from accepted findings, and I know of no positive evidence to support either.

The nature of the copy-text used in printing the play in 1623 Folio is central to deciding between these conflicting results. Most would agree that *Henry VIII* presents a more complex problem that of *Two Noble Kinsmen.* Hoy has maintained (and my results support his conclusions) that Fletcher revised scenes by Shakespeare; their hands do not appear to be so closely intermingled in *TNK.* A scientist without considerable textual expertise is not in the position to propose new hypotheses regarding copy-text. All I can do is present the results produced by this analysis of new evidence. Whether these findings will be dismissed by textual scholars or prompt a re-evaluation of the relation between the copy-text and all forms of internal evidence is uncertain.

8.4.3 Further Research and Other Applications of these Procedures

There are several areas in which further research could shed valuable light on the collaboration question in *Henry VIII* and *The Two Noble Kinsmen.* One regards the effect of collaboration on a writer's composition. An assumption that is implicit in almost every study of the authorship of these two plays is the validity of comparing an author's unaided composition to his share in a collaborative work. However, as noted in Chapter 7, Mincoff found changes in Fletcher's metrical characteristics when his unaided work is compared to several scenes attributed to him in collaborative works [105]. Of course, in order to test for such an alteration in Fletcher's use of function words, one would have to accept the attributions of traditional linguistic studies (such as Hoy's *Studies in Bibliography* series) to identify his share in collaborations.

The procedures developed in this dissertation are not suited to every type of authorship question in Elizabethan and Jacobean drama. By nature, discrimination techniques require that there be identifiable candidates for authorship who can be represented by a number of undisputed samples. Therefore, the discriminant analysis of function words could not easily be used to evaluate the integrity of the *Henry VI* plays or *Pericles.* Questions involving playwrights like

Christopher Marlowe and John Webster do not lend themselves to this approach, since only three plays by either author are suitable controls. Several questions surround plays associated with Ben Jonson. A detailed study of Jonson's works would be interesting in its own right. A large number of his plays (and other literary forms) have survived, and he gave a great deal of personal attention to their publication.

The so-called Beaumont and Fletcher canon contains a great many interesting authorship problems (which were the subject of Hoy's study). Many of these involve Fletcher and Phillip Massinger, who could be represented by the 15 plays agreed to be his unaided work. A critical old-spelling edition of these plays has been published in the last 15 years, so the Fletcher/Massinger collaborations would be a good choice for anyone wishing to test and refine the procedures presented in this study. (However, such an examination will probably not take place, since Massinger is less interesting than many other Jacobean playwrights.)

I feel that this study has been successful in achieving many of its goals. The effectiveness of approaching authorship questions through an analysis of function words has been demonstrated on a large number of undisputed samples by two authors. Function words may seem insignificant, but until recently no one has examined their pattern of occurrence in large samples of English text. Without computers such an examination is a gigantic undertaking, and without statistics it would be impossible to effectively use this information to assign authorship. As often noted in this dissertation, individual words do not discriminate well enough to be useful tests of authorship, but a multivariate analysis of a set of words can succeed even for short samples. This has allowed me to classify most of the individual disputed scenes in *TNK* and *H8* independently. For analyzing word rates, the kernel method is a statistically effective discriminant analysis procedure that allows for correlation between variables and non-normal data distributions.

Appendix A

The Sources and Printing of Early Editions

Questions concerning the reliability of the texts used in this investigation have arisen at several stages. In order to fully appreciate how this issue may affect the methods and results of this study, one must have a clear understanding of what is known about the process of writing and publication of dramas in the Elizabethan and Jacobean period. In particular, since in most cases the earliest versions of the plays are from printed quartos or folios, the reliability of the source (or sources) from which the publisher printed the play will have a tremendous influence on how much one can depend on the text to reproduce an author's intentions. A second major factor is the printing process itself, which could introduce a variety of alterations into the surviving versions of the plays. Results of research in these areas will not only influence which works one can assume to be suitable controls for comparison, but also help to determine what features of a text might be used to discriminate between authors.

A.1 The Sources

In many cases scholars have been able to deduce the nature of the sources used in printing from details in the plays themselves. Many of Shakespeare's plays exhibit features that cause scholars to postulate that the author's own papers lay behind the text. These could be "foul papers" resembling the *Sir Thomas More* fragment, untidy and carelessly written, with unclearly marked deletions, additions, revisions, and interlineations. Such drafts were then copied out neatly as "fair papers" for the theater company by either the author or a scribe. (It seems reasonable that an author would keep his foul papers after selling the fair copy to the company.) In either case the author's manuscript could include incomplete or vague stage directions, inconsistent speech prefixes or the presence of "ghost" characters. (This is discussed by Evans [34, pp. 228–230] and by Greg [43, pp. 106–114].) Thus in the 1599 Quarto of *Romeo and Juliet* (Q2), Juliet's mother is referred to in the speech prefixes as *Wife*, *Lady*, *Mother*, and *Lady Capulet*. An example of an indefinite stage direction is found in *Titus Andronicus*: "then enter ... and others as many as may be."

The inconsistencies and anomalies in either type of copy were certainly "foul" to the person responsible for directing the play's performances, so the author's papers were transcribed into a version which was submitted to the Master of Revels for examination and licensing. This copy was then used as the theater prompt-book. Again, certain features in a text (outlined by Greg [43, pp. 112-141]) tend to point towards prompt-book copy as the source of an edition. While in some cases the prompt-book may have been altered in the theater during production by the author himself, one must bear in mind that in all cases it is a version once removed (at least) from the author's draft and hence subject to errors of transcription. In addition, the prompt-book text may have been cut

due to censorship by the Master of Revels[1] editorial principle that published texts based on the author's foul or fair papers have more authority than prompt copy.

Naturally the theater companies had a different outlook concerning the relative values of their copies of a play. The prompt-book was of more use to their needs, and when they supplied a play to a printer they often happily parted with the manuscript sold to them by the author. When the author himself sold his play to a printer, it often appears that he sold the original foul papers, which he had kept after selling a fair copy to the company. (Bowers justifies this view [15, pp. 13–19].) Thus, an early printed edition of a play might be based upon the author's own papers (and may better reflect his personal habits) or upon a prompt-book modified for the theater.

Other forms of secondary copy besides the prompt-book are possible. Scribes were sometimes used to prepare special manuscript editions of a play for a gift to a patron or for publication. Other texts appear to have been memorially reconstructed by one or more actors, often for an unauthorized published edition. Six of the plays printed in the First Folio had been previously published in such an edition, known as a *bad quarto*. Where a more authoritative version exists, the bad quarto can be shown to include anticipation of text that actually occurs later in the play, recollection of earlier text, unconscious borrowings from other plays, and blatant ad-libbing. Such texts frequently contain vivid visual imagery from the earliest productions; for example, it is only in the bad quarto of *Hamlet* that we see the prince leaping into Ophelia's grave to confront Laertes. Finally, many plays were printed from some combination of these versions. New quartos were usually printed from previous editions, and several plays published in the First Folio were printed from such a quarto text collated against or conflated with another manuscript of more or less authority.

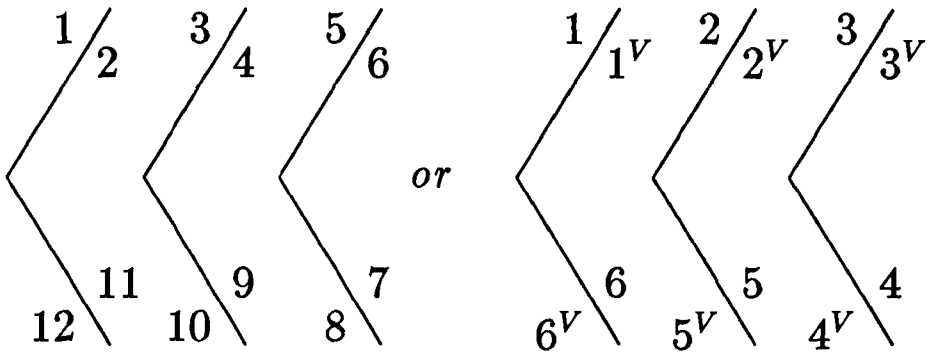
---

[1] Examples of small alterations in a text that might have resulted from censorship are given in Table 2–3 on page 48.

A.2 The Printing Process

In the last sixty years research in the field of bibliography has uncovered a wealth of information from early printed books regarding the printing process in Shakespeare's time, and many of these findings have had a great influence on editors of Renaissance drama. As usual, Shakespeare's texts initially received the most attention, and much of this was concentrated on the First Folio. Over two hundred copies of this first complete collection of his dramas have survived, and detailed examination of many of these copies has yielded valuable (and surprising) information about the treatment of the copy in the printing house. Some of the results regarding the Folio will be presented to show how the source texts could be altered in printing. Although in this study the specific details only apply to the sixteen plays of Shakespeare taken from the Folio, many of the general considerations apply equally well to texts printed as quartos or in the Beaumont and Fletcher folio.

First, the actual procedure used to divide the text into pages may have a great effect on the text. Until 1955 it was believed that the Folio was set by successive pages, that is, in the order that we read them today. Pages were grouped into *quires* before being bound in order to reduce both the amount of sewing required and the thickness of the book's back. If three sheets, or *formes,* are gathered together and then folded, the resulting book is "in sixes" and is composed of quires of six leaves or twelve pages. This was the most common arrangement in the early sixteenth century, and indeed the First Folio was printed in this manner. (The Beaumont and Fletcher 1647 Folio is "in fours.") If the pages were to be composed (set into type) in successive order, then the printing of the formes could not begin until seven pages had been set, as an examination of Figure A–1 shows. Not only would this require a large amount of type (often in short supply), the press(es) would have to stand idle while the compositor(s) set the initial pages of the quire.



A quire from a book "in sixes" is illustrated with two methods of identifying the pages. In the first the pages are numbered consecutively 1–12 and in the second by leaves (where the superscript "V" indicates that a page is the *verso* side of a leaf). The second method is more commonly used in bibliographical literature.

Figure A–1: A quire of six leaves

Hinman [49] has demonstrated conclusively that the First Folio was set "by formes:" the inner forme (pages 6 and 7) was set into type initially, then pages 5 and 8, etc. The main advantage in setting the First Folio in this way was the resulting effective balance between press-work and composition. The evidence indicates that usually two compositors worked simultaneously. Estimates of the time required to set a Folio forme and the time needed to print the required number of pages show that two compositors were required to keep the press operating at peak efficiency. The main implication of such a process is that the copy for the compositors, whether printed text or manuscript, must have been *cast off*, or divided into pages before composition could begin. Clearly if some slight miscalculation occurred the compositor might have found himself with too much or too little copy to fit onto pages 1 or 12 of the quire, which made up the last forme to be set. Padding out a page presented little difficulty. When faced

with too much copy, however, a compositor had several options. Text could be compressed if he altered the verse-lining; this usually necessitated abbreviating words or adjusting the spelling in order to fit two verses into a single line.

In some instances it is clear that text was actually deleted. The last page of the Folio text for *Much Ado about Nothing* is extremely crowded; *you* and *that* are contracted to $\overset{u}{y}$ and $\overset{t}{y}$, names are abbreviated, and a *tilde* is used to indicate the omission of an *n* after a vowel: *questiõ.* This text was printed from the good quarto of 1600, and comparison shows that the compositor was forced to omit a word in one speech and another entire line in setting this page, the first in quire L. Another crowded page that may have been so altered is the scene in *Antony and Cleopatra* in which Cleopatra surrenders to the Romans. Two consecutive short speeches are assigned to Proculeius, the first addressed to a Cleopatra at bay, and the next a command to his soldiers to guard the captured queen. Scholars have long recognized that at least a stage direction has vanished, and this cut is easily explained by an error in casting off the copy for the play.

The results of bibliography have thus disproved a long-standing assumption regarding the process employed in printing Shakespeare's First Folio. Since Hinman's findings were published it has become clear that more texts were set by formes than had been thought in McKerrow's day, including many quarto texts. (For example, parts of the quarto texts *Richard II, Richard III, Much Ado About Nothing* and *A Midsummer Night's Dream* were set from cast-off copy.) Detailed study of compositors' work has also led to a better understanding of how they might alter certain features of the text.

Editors have long assumed that the compositors who set type from manuscripts or printed text were tolerably faithful to their copy. Yet it has long been known that compositors felt quite comfortable in altering the copy spellings of common words to their own preferred forms (McKerrow provides examples [82]) and that they often varied spellings in order to justify lines. Recent studies have shown that they sometimes took greater liberties with their copy than this. A great deal of discussion regarding Shakespeare's use of contractions has not

clearly determined to what extent these forms reflect a playwright's intentions. While recognizing that these forms were intended to reflect the pronunciation of the play's lines, one must clearly consider the possibility that the author may not have been consistent in his usage. Examples in *Romeo and Juliet* and *Richard II* show that the printed form does not always agree with the metrical requirements of a line. Comparison of the Folio to quarto texts indicates that a theater scribe may have reproduced an actor's performance when copying a script. (Contractions and authorial intent are discussed more fully in Section 2.2.2 on page 31.) As always, the possibility exists that the orthography of the printed texts may reflect compositorial emendation.

The problems associated with the good quarto of *Hamlet* has led to a detailed examination of the work of the two compositors who set this text. Jenkins describes the sort of details that these same two workmen altered in printing Q2 of *Titus Andronicus* from the earlier good quarto [128, pp. 45–46]. They modernized some spellings (such as *whiles* to *whilst*), and corrected "mistakes" by making both the grammar and the metre more regular (for example, making the verb plural to agree with the noun, and contracting *overcome* to *oercome* and *the* to *th'*). In addition one of the pair goes so far as to set *you* for *ye* and *mine/thine* for *my/thy* before a vowel.[2]

The extensive comparisons between the quarto and Folio versions of several of Shakespeare's plays gives some indication of how well the Folio compositors reproduced their copy. Walker's evaluation of Compositor B's error rate in setting a little over half of the Folio text of *1 Henry IV* is not encouraging. Collation of the Folio text with the 1613 quarto of the text reveals 135 altered readings, of which only twenty-two are corrections of obvious errors in the quarto text. Thirty of the remaining 113 are deletions, twenty-eight are interpolations, and thirty-one involve altered or transposed individual words. On average he made

[2]Both the *-ine* and the *-y* forms of the possessive determiner were in free variation before vowels at this time, although the *-y* forms were spreading at the expense of the others. Barber discusses this change in Early Modern English [7, pp. 204–208].

some kind of error once every seventeen lines, a performance typical of his other work in the Folio [134, p. xviii]. The other compositor of this play, Compositor A, usually reproduced his copy more accurately, and indeed in this play his error rate was one-fifth that of B's. On the other hand, the work of the apprentice Compositor E (which was limited to certain of the tragedies that were set from quartos, including parts *Hamlet, King Lear* and *Othello*) was very much poorer, reflecting his inexperience and lack of mechanical skill. This apprentice's most infamous error occurs in *Hamlet* when, for the good quarto's reading "O treble woe," he substituted "Oh terrible woer."

Would alterations or mistakes on a compositor's part be corrected by a proof-reader? No doubt proof-reading practices would differ for every publication. But in the printing of the Shakespeare First Folio, a large and important new publication in its day, very little careful proof-reading was carried out and the corrections that were introduced usually corrupted the text further. Hinman describes the results of a collation of seventy-five copies of the First Folio [49], which reveal that what little proof-reading was done was based upon a principle of "intelligibility in a typographically neat page." About 750 of the Folio's 908 pages were examined and corrected during the printing process, and most of the press-variants found dealt with non-substantive readings such as turned letters, faulty space types, and other obvious typographical errors. Even such errors as these were often over-looked; for example, in two consecutive corrected pages of *Romeo and Juliet*, three obvious typographic errors were found and altered during printing, but nine other such errors in one of the pages and eight more in the other were not.

The small number of substantive press-variants (only a few dozen out of 500) do indicate that loyalty to the text was not the proof-reader's priority. Hinman describes two telling variants [49, pp. 240–243] from *1 Henry IV* that indicate that the printers would alter the text slightly rather than reset a large section of text. In one instance during a prose speech by Falstaff, the phrase "Sacke with lime in't" was set without the word *lime*. To add the word without rearranging

the lineation of the entire passage would have been impossible, so the compositor captured the sense of the passage by a simple substitution, correcting to "Sack with lime". He could have made this emendation without referring to the copy text, since "Lime in this Sacke too" appears in the immediate context. Near the end of the play, Falstaff (again in a prose passage) hopes to be rewarded for his part in the battle; the quarto text reads:

> *Fal.*
> Ile follow as they say for reward. Hee that rewardes mee God reward him. If I do growe great, ile growe lesse, for ile purge and leaue Sacke, and liue cleanlie as a noble man should do. *Exit.*

In setting the Folio text from the quarto, the compositor repeated the word *great.*

> *Fal.* Ile follow as they say, for Reward. Hee that rewards me, heauen reward him. If I do grow great great, Ile grow lesse? For Ile purge, and leaue Sacke, and liue cleanly, as a Nobleman should do. *Exit.*

Simply removing the word would have left a typographical flaw since prose lines were always justified, so someone simply thought of a word of the same length that produced a sensible statement: in this case, *again.* In fact, Falstaff had never been great in the sense intended, *noble.* In any case the Folio's reading "If I do grow great again" is a corruption of the copy text. While a few examples in the Folio do indicate that the copy was consulted in correction (two missing lines in *Richard II* were reinstated), the evidence clearly shows that for the most part the Folio proof-reader's corrections are not authoritative.

In the absence of effective proof-reading, the degree to which the resulting printed text faithfully reproduces the copy therefore depends on the integrity and initial skill of the compositors themselves. Both skill and integrity have been shown to have been lacking sometimes, and therefore in some details the

published text may not exactly reproduce the copy which the printer received from the author or company. Moreover this copy was sometimes once removed (or further) from the author's own manuscripts. Thus certain features of the author's hand will clearly have been altered or obscured in the only texts that have survived. Any researcher who does not recognize these problems and attempt to make allowances for them in his methods works in truly blissful ignorance.

Appendix B

Frequency Distributions of Marker Words

In *Inference and Disputed Authorship* Mosteller and Wallace demonstrate that the negative binomial describes the distribution of word frequencies in 200 word samples of Hamilton and Madison much better than the Poisson distribution [113]. In an authorship study of the Middle English *Pearl* poems, McColly and Weier analyze function word frequencies using a likelihood-ratio approach [80]. This analysis assumes that these frequencies are distributed according to the Poisson distribution, although the authors note that there are some objections to this assumption. They do not present goodness-of-fit results for word frequencies in the works they analyze but state that the Poisson is "only an approximation" and cite its computational tractability. In "The Use of Function Word Frequencies as Indicators of Style" [26] Damerau defines context-free words to be those that can be described by the Poisson distribution.

In Chapter 6 the Kolomogorov goodness-of-fit test was used to show that word *rates* were not distributed normally when measured in scenes of at least 1000 words in the 26 plays in the control set. One should note that the distributions describing rate of occurrence and frequency of occurrence of the same word may have different forms. The analyses of Chapters 6 and 7 use rates in order to facilitate the comparison of samples of different length. This appendix examines

the frequency distribution for *counts* of the 14 individual marker words selected in Chapter 5.

Each of the words was counted in blocks of 250 words: a total of 530 blocks taken from the 6 Fletcher plays in the control set and 1694 from the 20 Shakespeare plays. For each word in each author, a table lists the number of blocks containing n occurrences is listed together with the expected number of occurrences for the Poisson and the negative binomial distributions. (These tables begin on page 361). Below these values, the mean number of occurrences per 250 word block is listed, together with the standard error of this mean, the variance and the standard deviation. The results of the χ^2 goodness-of-fit test for each distribution follow.

The results indicate that the pattern of occurrence found by Mosteller and Wallace in 18th century American prose is also evident in English Renaissance drama. At the 5% level of significance the Poisson only satisfactorily fits the distributions for *in* and *these* in the Fletcher samples and none of Shakespeare's word distributions. On the other hand, the negative binomial accurately describes the pattern of occurrence for all 14 words in Shakespeare and only fails to fit the occurrences of *these* in Fletcher.

A number of the χ^2 values for the Poisson are astronomical. The overall unsuitability of the Poisson model raises further suspicions about the validity of McColly and Weier's statistical analysis. In addition, if Damerau's definition of contextuality is accepted, almost all of these words depend on context. These results and Mosteller and Wallace's findings suggest that the usefulness of the Poisson distribution for studying either authorship or contextuality is extremely limited (at best). These results also imply that a univariate parametric discriminant analysis approach following Mosteller and Wallace could be based on this data. However, the degree of correlation described in Chapter 5 indicates that a true multivariate approach (such as the distribution-free procedures outline in Chapter 6) is desirable.

B.1 Interpretation of the Negative Binomial

In his article "Fitting the Negative Binomial Distribution to Biological Data" [13], Bliss discusses several underlying models for the negative binomial. The Poisson distribution assumes that the number of observations occurring in a sequence of repeated observations has a constant expectation of occurrence. If this is not the case, then the data may be represented by a mixture of several Poisson distributions. In this situation the means can represent a continuous variable, and if they are distributed according to certain distributions then the original data will be described by the negative binomial.

Mosteller and Wallace also follow this explanation, asserting that the occurrence of word counts is an example of *contagious* distributions. [113, pp. 93–95]. Contagion is used to describe situations where the occurrence of one individual increases the chance of another individual in the same unit of observation. Bliss notes that the negative binomial can be used to describe the occurrence of bacteria in a milk film and that it figures prominently in accident statistics. He also states that, for many populations, agreement with the Poisson at low densities may be accompanied by agreement with the negative binomial for higher densities.

B.2 Details of the Calculations

Calculation of the expected numbers for the Poisson is very straightforward and requires only one parameter, the mean rate of occurrence μ. Estimating this from the sample mean m, the probability that a block will have x occurrences is:

$$\frac{e^{-m}\, m^x}{x!}$$

For the χ^2 test, counts with low expected frequencies were amalgamated so that the expected number was greater than 1.0. Since one parameter was estimated

from the sample data, the degrees for freedom is $n - 2$, where n is the number of cells remaining after amalgamations [151, pp. 76–77]

The negative binomial is specified by two parameters. The first is the mean, which is again estimated by m. The second is a positive exponent k, and from the distribution's expression

$$(q - p)^{-k}$$

its similarity to the binomial is apparent. Several methods of estimating k have been proposed. The simplest is the moment solution $\hat{k} = m^2/(s^2 - m)$, where s^2 is the sample variance. Bliss discusses the efficiency of this estimate for various values of m and demonstrates a better estimation using a maximum likelihood approach set forth by Fisher [13]. This method was implemented in a Pascal program and used in the goodness-of-fit tests that follow.

Once k has been estimated, the recursion formula for calculating the expected frequencies $f(x)$ for blocks containing x occurrences is:

$$f(0) = N/q^k$$
$$f(x) = \frac{(k + x - 1)}{x} \frac{p}{q} f(x - 1)$$

where N is the total number of blocks and $p = q - 1 = m/k$. Before calculating χ^2, counts were amalgamated until the expected frequency was at least 1.0. Since two parameters were estimated from the sample data, the degrees of freedom used in testing the negative binomial is $n - 3$.

ALL

Fletcher

#	Obs.	Poi.	N.B.
0	116	79.27	115.84
1	145	150.62	143.17
2	115	143.08	113.50
3	61	90.62	73.20
4	47	43.04	41.81
5	28	16.36	22.02
6	10	5.18	10.95
7	6	1.41	5.21
8	0	0.33	2.40
9	1	0.07	1.08
10	1	0.01	0.47

Mean number of occurrences per block = 1.9000; SE = 0.0738
Variance = 2.8879 Standard Deviation = 1.6994

Poisson Distribution:
Chi-square = 66.47 df = 6 prob = 2.18E-12
Negative Binomial Distribution:
p = 0.5372 k = 3.5365 se for k = 0.6771
Chi-square = 7.08 df = 7 prob = 4.21E-01

Shakespeare

#	Obs.	Poi.	N.B.
0	632	574.77	632.97
1	571	621.26	569.12
2	303	335.75	303.70
3	126	120.97	125.07
4	42	32.69	43.89
5	15	7.07	13.80
6	2	1.27	4.00
7	3	0.20	1.09

Mean number of occurrences per block = 1.0809; SE = 0.0277
Variance = 1.3018 Standard Deviation = 1.1410

Poisson Distribution:
Chi-square = 33.21 df = 5 prob = 3.42E-06
Negative Binomial Distribution:
p = 0.2021 k = 5.3471 se for k = 1.1690
Chi-square = 4.55 df = 5 prob = 4.74E-01

ARE

Fletcher

#	Obs.	Poi.	N.B.
0	126	102.46	132.03
1	164	168.38	156.57
2	120	138.36	114.63
3	67	75.79	66.59
4	26	31.14	33.65
5	13	10.24	15.47
6	6	2.80	6.65
7	5	0.66	2.71
8	2	0.14	1.06
9	0	0.02	0.40
10	1	0.00	0.15

Mean number of occurrences per block = 1.6434; SE = 0.0667
Variance = 2.3546 Standard Deviation = 1.5345

Poisson Distribution:
Chi-square = 40.26 df = 5 prob = 1.32E-07
Negative Binomial Distribution:
p = 0.3858 k = 4.2594 se for k = 0.9799
Chi-square = 6.19 df = 6 prob = 4.02E-01

Shakespeare

#	Obs.	Poi.	N.B.
0	721	595.14	715.43
1	498	622.55	514.15
2	278	325.61	265.21
3	117	113.53	118.87
4	47	29.69	49.26
5	22	6.21	19.41
6	9	1.08	7.39
7	1	0.16	2.74
8	1	0.02	1.00

Mean number of occurrences per block = 1.0460; SE = 0.0298
Variance = 1.5053 Standard Deviation = 1.2269

Poisson Distribution:
Chi-square = 183.67 df = 5 prob = 6.82E-14
Negative Binomial Distribution:
p = 0.4555 k = 2.2962 se for k = 0.2891
Chi-square = 2.80 df = 5 prob = 7.30E-01

DARE

Fletcher

#	Obs.	Poi.	N.B.
0	403	385.29	402.58
1	94	122.86	96.13
2	27	19.59	23.55
3	3	2.08	5.82
4	3	0.17	1.44

Mean number of occurrences per block = 0.3189; SE = 0.0282
Variance = 0.4218 Standard Deviation = 0.6494

Poisson Distribution:
Chi-square = 16.66 df = 2 prob = 2.41E-04
Negative Binomial Distribution:
p = 0.3353 k = 0.9509 se for k = 0.3156
Chi-square = 3.60 df = 2 prob = 1.65E-01

Shakespeare

#	Obs.	Poi.	N.B.
0	1601	1591.25	1601.12
1	83	99.57	81.65
2	7	3.12	9.64
3	3	0.06	1.35

Mean number of occurrences per block = 0.0626; SE = 0.0068
Variance = 0.0776 Standard Deviation = 0.2786

Poisson Distribution:
Chi-square = 17.44 df = 1 prob = 2.96E-05
Negative Binomial Distribution:
p = 0.2270 k = 0.2756 se for k = 0.1033
Chi-square = 2.75 df = 1 prob = 9.72E-02

DID

Fletcher

#	Obs.	Poi.	N.B.
0	447	433.11	446.28
1	62	87.44	65.88
2	18	8.83	13.72
3	3	0.59	3.13

Mean number of occurrences per block = 0.2019; SE = 0.0223
Variance = 0.2635 Standard Deviation = 0.5133

Poisson Distribution:
Chi-square = 22.08 df = 1 prob = 2.61E-06
Negative Binomial Distribution:
p = 0.3676 k = 0.5492 se for k = 0.2000
Chi-square = 1.57 df = 1 prob = 2.10E-01

Shakespeare

#	Obs.	Poi.	N.B.
0	1136	1006.45	1140.93
1	368	524.02	350.36
2	116	136.42	125.65
3	44	23.68	47.22
4	15	3.08	18.15
5	7	0.32	7.07
6	5	0.03	2.78
7	2	0.00	1.10
8	0	0.00	0.44
9	0	0.00	0.17
10	0	0.00	0.07
11	1	0.00	0.03

Mean number of occurrences per block = 0.5207; SE = 0.0235
Variance = 0.9349 Standard Deviation = 0.9669

Poisson Distribution:
Chi-square = 289.24 df = 3 prob = 8.93E-14
Negative Binomial Distribution:
p = 0.6955 k = 0.7486 se for k = 0.0838
Chi-square = 4.98 df = 5 prob = 4.18E-01

IN

Fletcher

#	Obs.	Poi.	N.B.
0	60	62.85	68.42
1	143	134.01	134.57
2	144	142.86	137.55
3	98	101.53	97.29
4	45	54.12	53.49
5	27	23.08	24.36
6	8	8.20	9.56
7	2	2.50	3.32
8	1	0.67	1.04
9	1	0.16	0.30
10	0	0.03	0.08
11	1	0.01	0.02

Mean number of occurrences per block = 2.1321; SE = 0.0665
Variance = 2.3417 Standard Deviation = 1.5303

Poisson Distribution:
Chi-square = 3.87 df = 6 prob = 6.94E-01
Negative Binomial Distribution:
p = 0.0840 k = 25.3818 se for k = 19.3436
Chi-square = 5.97 df = 6 prob = 4.26E-01

Shakespeare

#	Obs.	Poi.	N.B.
0	83	57.53	85.53
1	218	194.60	226.57
2	340	329.12	324.67
3	352	371.08	333.63
4	261	313.80	275.21
5	179	212.29	193.55
6	107	119.68	120.43
7	87	57.83	67.96
8	37	24.45	35.40
9	17	9.19	17.24
10	7	3.11	7.93
11	5	0.96	3.47
12	1	0.27	1.46

Mean number of occurrences per block = 3.3825; SE = 0.0505
Variance = 4.3261 Standard Deviation = 2.0799

Poisson Distribution:
Chi-square = 82.15 df = 10 prob = 2.20E-13
Negative Binomial Distribution:
p = 0.2768 k = 12.2183 se for k = 1.9851
Chi-square = 11.80 df = 10 prob = 3.009E-01

MUST

Fletcher

#	Obs	Poi.	N.B.
0	268	237.25	270.17
1	161	190.69	154.12
2	58	76.64	66.33
3	31	20.53	25.45
4	10	4.13	9.17
5	0	0.66	3.18
6	0	0.09	1.07
7	1	0.01	0.35
8	0	0.00	0.11
9	1	0.00	0.04

Mean number of occurrences per block = 0.8038; SE = 0.0470
Variance = 1.1713 Standard Deviation = 1.0822

Poisson Distribution:
Chi-square = 28.82 df = 3 prob = 2.44E-06
Negative Binomial Distribution:
p = 0.4090 k = 1.9653 se for k = 0.4760
Chi-square = 5.95 df = 4 prob = 2.03E-01

Shakespeare

#	Obs.	Poi.	N.B.
0	1144	1068.91	1143.49
1	385	492.18	386.64
2	118	113.31	116.72
3	34	17.39	33.83
4	8	2.00	9.60
5	5	0.18	2.69

Mean number of occurrences per block = 0.4604; SE = 0.0192
Variance = 0.6242 Standard Deviation = 0.7901

Poisson Distribution:
Chi-square = 98.15 df = 3 prob = 3.69E-14
Negative Binomial Distribution:
p = 0.3618 k = 1.2727 se for k = 0.2078
Chi-square = 2.27 df = 3 prob = 5.18E-01

NO

Fletcher

#	Obs.	Poi.	N.B.
0	126	108.63	128.08
1	175	172.17	163.63
2	106	136.44	120.39
3	67	72.08	66.83
4	34	28.56	31.07
5	17	9.05	12.76
6	4	2.39	4.78
7	1	0.54	1.67

Mean number of occurrences per block = 1.5849; SE = 0.0607
Variance = 1.9521 Standard Deviation = 1.3972

Poisson Distribution:
Chi-square = 19.44 df = 5 prob = 1.59E-03
Negative Binomial Distribution:
p = 0.2406 k = 6.5878 se for k = 2.2282
Chi-square = 4.63 df = 5 prob = 4.63E-01

Shakespeare

#	Obs.	Poi.	N.B.
0	619	533.89	617.72
1	539	616.46	547.62
2	318	355.90	306.32
3	137	136.98	137.95
4	55	39.54	54.60
5	14	9.13	19.82
6	8	1.76	6.77
7	2	0.29	2.20
8	1	0.04	0.69
9	0	0.01	0.21
10	1	0.00	0.06

Mean number of occurrences per block = 1.1547; SE = 0.0299
Variance = 1.5153 Standard Deviation = 1.2310

Poisson Distribution:
Chi-square = 82.81 df = 5 prob = 3.09E-14
Negative Binomial Distribution:
p = 0.3025 k = 3.8174 se for k = 0.6146
Chi-square = 2.75 df = 5 prob = 7.39E-01

NOW

Fletcher

#	Obs.	Poi.	N.B.
0	142	123.97	149.83
1	173	180.11	165.63
2	125	130.83	111.35
3	50	63.36	58.78
4	21	23.01	26.79
5	12	6.69	11.05
6	4	1.62	4.24
7	0	0.34	1.54
8	1	0.06	0.53
9	1	0.01	0.18
10	0	0.00	0.06
11	0	0.00	0.02
12	1	0.00	0.01

Mean number of occurrences per block = 1.4528; SE = 0.0619
Variance = 2.0290 Standard Deviation = 1.4244

Poisson Distribution:
Chi-square = 22.57 df = 5 prob = 4.08E-04
Negative Binomial Distribution:
p = 0.3142 k = 4.6234 se for k = 1.2060
Chi-square = 5.26 df = 5 prob = 3.85E-01

Shakespeare

#	Obs.	Poi.	N.B.
0	837	761.70	838.73
1	523	608.82	521.28
2	224	243.31	219.96
3	74	64.83	78.19
4	27	12.95	25.19
5	5	2.07	7.61
6	1	0.28	2.20
7	2	0.03	0.61
8	1	0.00	0.17

Mean number of occurrences per block = 0.7993; SE = 0.0248
Variance = 1.0383 Standard Deviation = 1.0189

Poisson Distribution:
Chi-square = 56.00 df = 4 prob = 2.01E-11
Negative Binomial Distribution:
p = 0.2860 k = 2.7943 se for k = 0.4830
Chi-square = 1.68 df = 4 prob = 7.94E-01

OF

Fletcher

#	Obs.	Poi.	N.B.
0	34	21.12	31.77
1	74	68.06	78.54
2	107	109.67	106.23
3	107	117.81	104.03
4	80	94.92	82.47
5	58	61.18	56.14
6	33	32.86	34.02
7	20	15.13	18.81
8	9	6.09	9.64
9	4	2.18	4.65
10	2	0.70	2.12
11	1	0.21	0.93
12	1	0.06	0.39

Mean number of occurrences per block = 3.2226; SE = 0.0891
Variance = 4.2036 Standard Deviation = 2.0503

Poisson Distribution:
Chi-square = 22.38 df = 8 prob = 4.26E-03
Negative Binomial Distribution:
p = 0.3035 k = 10.6167 se for k = 2.9008
Chi-square = 1.25 df = 9 prob = 9.99E-01

OF (continued)

Shakespeare

#	Obs.	Poi.	N.B.
0	49	11.65	34.88
1	100	57.99	107.63
2	175	144.40	186.53
3	216	239.70	239.18
4	265	298.43	252.74
5	247	297.23	232.89
6	199	246.70	193.59
7	146	175.50	148.46
8	121	109.25	106.67
9	65	60.45	72.64
10	48	30.10	47.28
11	27	13.63	29.61
12	12	5.66	17.94
13	6	2.17	10.55
14	11	0.77	6.05
15	2	0.26	3.39
16	2	0.08	1.86
17	2	0.02	1.01
18	0	0.01	0.53
19	0	0.00	0.28
20	0	0.00	0.14
21	0	0.00	0.07
22	0	0.00	0.04
23	0	0.00	0.02
24	0	0.00	0.01
25	1	0.00	0.00

Mean number of occurrences per block = 4.9799; SE = 0.0689
Variance = 8.0362 Standard Deviation = 2.8348

Poisson Distribution:
Chi-square = 474.55 df = 13 prob = 0.0
Negative Binomial Distribution:
p = 0.6137 k = 8.1144 se for k = 0.7626
Chi-square = 23.37 df = 16 prob = 1.04E-01

SURE

Fletcher

#	Obs.	Poi.	N.B.
0	362	353.26	362.34
1	130	143.31	129.35
2	31	29.07	30.85
3	6	3.93	6.14
4	0	0.40	1.10
5	1	0.03	0.18

Mean number of occurrences per block = 0.4057; SE = 0.0296
Variance = 0.4646 Standard Deviation = 0.6816

Poisson Distribution:
Chi-square = 3.18 df = 2 prob = 2.04E-01
Negative Binomial Distribution:
p = 0.1364 k = 2.9744 se for k = 1.7117
Chi-square = 0.07 df = 2 prob = 9.65E-01

Shakespeare

#	Obs.	Poi.	N.B.
0	1554	1547.71	1553.95
1	128	139.79	128.20
2	11	6.31	10.84
3	1	0.19	0.92

Mean number of occurrences per block = 0.0903; SE = 0.0076
Variance = 0.0987 Standard Deviation = 0.3142

Poisson Distribution:
Chi-square = 5.67 df = 1 prob = 1.73E-02
Negative Binomial Distribution:
p = 0.0948 k = 0.9532 se for k = 0.5325
Chi-square = 0.01 df = 0

THE

Fletcher

#	Obs.	Poi.	N.B.
0	8	1.46	6.02
1	17	8.61	20.95
2	35	25.37	40.76
3	60	49.84	58.43
4	77	73.44	68.79
5	69	86.57	70.42
6	78	85.05	64.89
7	48	71.62	55.04
8	41	52.77	43.67
9	32	34.56	32.78
10	21	20.37	23.49
11	12	10.92	16.18
12	10	5.36	10.76
13	8	2.43	6.95
14	4	1.02	4.37
15	4	0.40	2.68
16	1	0.15	1.61
17	4	0.05	0.95
18	1	0.02	0.55

Mean number of occurrences per block = 5.8943; SE = 0.1385
Variance = 10.1703 Standard Deviation = 3.1891

Poisson Distribution:
Chi-square = 168.04 df = 13 prob = 6.82E-14
Negative Binomial Distribution:
p = 0.6931 k = 8.5042 se for k = 1.2989
Chi-square = 17.58 df = 15 prob = 2.85E-01

THE (continued)

Shakespeare

#	Obs.	Poi.	N.B.
0	5	0.49	5.74
1	28	4.02	23.69
2	40	16.37	54.75
3	94	44.41	93.35
4	133	90.38	130.86
5	174	147.13	159.65
6	165	199.61	175.43
7	181	232.11	177.57
8	172	236.17	168.21
9	159	213.60	150.85
10	145	173.87	129.18
11	104	128.66	106.36
12	70	87.28	84.63
13	61	54.65	65.38
14	50	31.77	49.19
15	28	17.24	36.16
16	18	8.77	26.04
17	19	4.20	18.40
18	12	1.90	12.78
19	12	0.81	8.74
20	7	0.33	5.90
21	7	0.13	3.93
22	1	0.05	2.58
23	5	0.02	1.68
24	0	0.01	1.08
25	0	0.00	0.69
26	2	0.00	0.44
27	1	0.00	0.27
28	1	0.00	0.17

Mean number of occurrences per block = 8.1399; SE = 0.0983
Variance = 16.3850 Standard Deviation = 4.0478

Poisson Distribution:
Chi-square = 1385.87 df = 18 prob = 0.0
Negative Binomial Distribution:
p = 0.9713 k = 8.3808 se for k = 0.5928
Chi-square = 32.77 df = 23 prob = 8.52E-02

THESE

Fletcher

#	Obs.	Poi.	N.B.
0	289	248.24	292.24
1	154	188.29	139.03
2	41	71.41	58.99
3	27	18.05	24.01
4	11	3.42	9.57
5	7	0.52	3.76
6	1	0.07	1.47

Mean number of occurrences per block = 0.7585; SE = 0.0479
Variance = 1.2157 Standard Deviation = 1.1026

Poisson Distribution:
Chi-square = 86.39 df = 3 prob = 2.82E-14
Negative Binomial Distribution:
p = 0.5943 k = 1.2763 se for k = 0.2533
Chi-square = 10.65 df = 4 prob = 3.08E-02

Shakespeare

#	Obs.	Poi.	N.B.
0	1220	1152.81	1219.51
1	341	443.70	344.93
2	103	85.39	94.51
3	19	10.95	25.61
4	9	1.05	6.90
5	1	0.08	1.86
6	0	0.01	0.50
7	1	0.00	0.13

Mean number of occurrences per block = 0.3849; SE = 0.0176
Variance = 0.5263 Standard Deviation = 0.7255

Poisson Distribution:
Chi-square = 122.44 df = 3 prob = 4.55E-14
Negative Binomial Distribution:
p = 0.3608 k = 1.0668 se for k = 0.1808
Chi-square = 3.25 df = 3 prob = 3.55E-01

TOO

Fletcher

#	Obs.	Poi.	N.B.
0	198	174.77	197.55
1	171	193.89	174.19
2	98	107.55	94.67
3	44	39.77	40.78
4	12	11.03	15.27
5	3	2.45	5.20
6	3	0.45	1.65
7	0	0.07	0.50
8	1	0.01	0.14

Mean number of occurrences per block = 1.1094; SE = 0.0516
Variance = 1.4096 Standard Deviation = 1.1872

Poisson Distribution:
Chi-square = 12.59 df = 4 prob = 1.35E-02
Negative Binomial Distribution:
p = 0.2583 k = 4.2959 se for k = 1.3870
Chi-square = 3.32 df = 4 prob = 5.05E-01

Shakespeare

#	Obs.	Poi.	N.B.
0	1187	1117.30	1187.36
1	363	464.99	363.14
2	107	96.76	103.67
3	27	13.42	28.89
4	4	1.40	7.95
5	5	0.12	2.17
6	1	0.01	0.59

Mean number of occurrences per block = 0.4162; SE = 0.0183
Variance = 0.5703 Standard Deviation = 0.7552

Poisson Distribution:
Chi-square = 88.81 df = 3 prob = 3.75E-14
Negative Binomial Distribution:
p = 0.3608 k = 1.1535 se for k = 0.1915
Chi-square = 5.98 df = 3 prob = 1.12E-01

WHICH

Fletcher

#	Obs.	Poi.	N.B.
0	411	388.95	410.45
1	86	120.35	87.67
2	22	18.62	22.94
3	10	1.92	6.37
4	1	0.15	1.82

Mean number of occurrences per block = 0.3094; SE = 0.0286
Variance = 0.4334 Standard Deviation = 0.6583

Poisson Distribution:
Chi-square = 50.22 df = 2 prob = 1.25E-11
Negative Binomial Distribution:
p = 0.4487 k = 0.6896 se for k = 0.1986
Chi-square = 2.51 df = 2 prob = 2.85E-01

Shakespeare

#	Obs.	Poi.	N.B.
0	929	825.38	924.81
1	462	593.45	475.63
2	202	213.35	190.01
3	67	51.13	68.64
4	21	9.19	23.48
5	11	1.32	7.76
6	2	0.16	2.51

Mean number of occurrences per block = 0.7190; SE = 0.0242
Variance = 0.9925 Standard Deviation = 0.9962

Poisson Distribution:
Chi-square = 152.49 df = 4 prob = 6.11E-14
Negative Binomial Distribution:
p = 0.3980 k = 1.8064 se for k = 0.2600
Chi-square = 2.92 df = 4 prob = 5.72E-01

Appendix C

Translation List for Spelling Variants

This appendix contains the translation list used with program REPLACE to handle variant spellings and homonyms for some common words. The basic approach to replacement was described in Section 2.5.3 (beginning on page 60). Each entry in the translation list is of the pattern: original form found in the text file, followed by an equals sign (=), followed by the new string to replace the original. For example:

```
doeth=doth
```

indicates that the variant form *doeth* will be standardized to *doth,* and:

```
off#1=of
```

shows how the hash suffix is used to mark instances of *off* that would be modernized to *of.*

Another list precedes the software translation list. It was compiled as an information list when the translations were first developed. It is not intended to be used with REPLACE, since a "definition" is often given instead of a replacement string for forms that are not altered by the program. For example, `bee#1` is used to indicate occurrences referring to the insect. This form and others like it remain in the version of the text produced by REPLACE. This list also indicates that occurrences of *and* as a subordinator meaning *and if* are marked `and#1`. A number of the codings found in the information list indicate how certain unique orthographical or linguistic forms were distinguished from the ordinary usage. These include the occurrence of "No had" in TLN 1932 of *King John,* meaning "Had I not," and the Folio reading "Are" in TLN 1367 of *The Taming of the Shrew* to indicate the notes of the scale "A Re." To distinguish the entries in this information list from the translations, the equal sign has been replaced with a dash and the normal font has been used for the "definition." (A number of

the entries in the information list duplicate entries in the software replacement list without providing any more information.)

In the actual translation list, a number of entries are set up for two-step replacement in conjunction with the contraction expansion list (provided in the next Appendix). For example, `'has#1` in the input text file is replaced with the string `h'as#1` at the variant translation stage; when contractions are expanded, this form is replaced with `he has`. In addition, the list reflects some codings that were not used in the final versions of the text files. These include `an#3`, which was originally intended to mark occurrences of *an other,* and `to#2`, which was used at first to mark occurrences of *to day etc.* (Recall from Section 2.5 that these forms are joined using the underscore character in the final versions of the texts.)

The entries in the list were compiled from an examination of the 34 plays used in this study. Many of the forms that are marked in my text files reflect the original markings found in the files I obtained from the Oxford University Press Shakespeare Department. (These include the coding `by#1` for both occurrences in the phrase "by and by," and the distinction of *a* in occurrences of `a#2 while`. Obviously the retention of these codings will affect the counts for some function words, but the numbers involved are fairly small.) If another text not in this set was to be processed using this list and REPLACE, one would have to examine that text carefully in order to mark words according to this coding scheme. New translations would probably have to be added. (These comments also apply to the contraction expansion list.)

Program REPLACE is written in the Pascal programming language, using the approach outlined in *Software Tools in Pascal,* by Brian W. Kernighan and P. J. Plauger (Addison-Wesley, 1981). It should be portable, if one can write the dozen or so primitive input/output procedures and functions described in the book. Currently I have versions of the program that run under DEC's VAX VMS operating system and on UNIX systems with the Berkeley Pascal compiler. I hope to implement the software on the IBM Personal Computer in the near future. The software and the lists are freely available. They have been deposited in the Oxford University Computing Service's Text Archive (13 Banbury Rd., Oxford OX2 6NN, England). To enquire about any software or data, please contact me through the following address:

103 Darwin Rd., Oak Ridge, TN 37830 USA

or contact my supervisor, Prof. Sidney Michaelson, at:

The Department of Computer Science,
JCMB, The King's Buildings, Mayfield Rd.,
Edinburgh EH9 3JZ, Scotland.

Information List:

&#1 — and#1
'has#1 — he has
a#1 — he
a#2 — [prep., eg "a horseback"]
a#3 — on
a#4 — ah
a#5 — of
a#6 — [other]
a#7 — have
a#8 — in
an#1 — and if
an#2 — on
an#3 — [eg another]
an#4 — Anne
and#1 — and if
Are#1 — A Re [Shrew, TLN 1367]
art#1 — [noun]
art'#1 — art
Arte#1 — art#1
at'#1 — at the
bad#1 — bade
be#1 — by
bee#1 — [noun]
bee#2 — by
Been#1 — [Latin]
being#1 — [noun]
bene#1 — [Latin]
but#1 — butt [cask]
but#2 — butt [verb]
but#3 — butt [buttocks]
by#1 — [in "by and by"]
by#2 — buy
could#1 — cold
de#1 — do
ayle#1 — i will
could#1 — cold
deer#1 — dear
di'd#1 — died
di'de#1 — died
dide#1 — died
die#1 — [sing. of "dice"]
Doe#1 — [a deer, a female deer]
don#1 — done
dost#1 — does it
E'n#1 — even
fort#1 — for it
Gives#1 — gyves [noun]
h'as#1 — he has
ha#1 — have
ha'#1 — have
ha's#9 — he has
had#1 — ["No had" = "had i not"]
has#1 — he has
has't#2 — hast thou
hast#1 — haste
hauing#1 — [noun]
heard#1 — herd
heele#1 — heel [noun]
heere#1 — hear
hel#1 — hell
hell#1 — hell
here#1 — hear
i#1 — aye
il#1 — ill [noun]
Ile#1 — isle
ill#1 — ill [noun]
in#1 — e'en
it#1 — its
lets#1 — [noun]
Maie#1 — May [month]
May#1 — May [month]
might#1 — [noun]
mine#1 — [noun]
no#1 — not
not#1 — knot
o#1 — of
o#2 — he
o#3 — [other]
o#4 — [Latin]
o#5 — [o'clock]
o'#1 — of
o'#2 — he
o'#3 — [other]
o'#4 — [Latin]
o'#5 — [o'clock]
of#1 — off
off#1 — of
on#1 — one

one#1 — on
or#1 — our
oth#1 — oath
our#1 — ours
shee#1 — [noun]
she#1 — [noun]
so#1 — [so, so]
the#1 — thee
there#1 — their
to#1 — too
to#2 — [eg today,tomorrow]
to#3 — two
too#1 — to
too#2 — [eg today,tommorrow]
too#3 — two
wast#1 — waste
we#1 — oui
wee#1 — oui
well#1 — [noun]
were#1 — wear
wert#1 — were it
whe'r#1 — whether
where#1 — whether
where#2 — wherever
wil#1 — will [noun]
wild#1 — willed
will#1 — [noun]
wilt#1 — will it
yare#1 — "quick"
your#1 — you are
your'#1 — you are

Software Replacement List:

```
&=and
&#1=and
&c=etc
'a#1=he
'a#3=on
'beseech=beseech
'blesse=bless
'fore=before
'gainst=against
'had=had
'had#1=had#1
'has=has
'has#1=h'as#1
'haue=haue
'is=he's
'm='em
'mongst=amongst
'pre-thee=prithee
'preethe=prithee
'pree-thee=prithee
'preethee=prithee
'prethe=prithee
'prethee=prithee
'prythee=prithee
'saue=saue
't=it
't'had='thad
'ts='tis
'tweene=between
'twer='twere
'twil='twill
'twold='twould
'twou'd='twould
'ye=ye
a#1=he
a#3=on
a#4=ah
a#5=of
a#7=haue
a#8=in
a'#1=he
a'#3=on
a'#4=ah
a'#5=of
a'#7=haue
a'#8=in
a'th=o'th'
a'th'=o'th'
al=all
alls=all's
alreadie=already
an#1=and
an#2=on
an#4=Anne
an#5=and
an'#5=and
an't=and't#1
and#1=and
angrie=angry
anie=any
armie=army
Arte#1=art#1
at'=at'th'
ath=o'th'
ath'=at'th'
att=at
awaie=away
ayle#1=i'll
bad#1=bade
be#1=by
be'st=beest
beautie=beauty
bee=be
bee#2=by
bee'st=beest
beeing=being
beene=been
bene=been
beutie=beauty
bewtie=beauty
bi'th'=by'th'
bin=been
bith'=by'th'
bloodie=bloody
bloudie=bloody
bodie=body
bonnie=bonny
bountie=bounty
```

```
brauerie=brauery
burie=bury
busie=busy
by#2=buy
by'th=by'th'
byn=been
byth'=by'th'
carrie=carry
ceremonie=ceremony
charitie=charity
citie=city
companie=company
contrarie=contrary
controuersie=controuersy
could#1=cold
countie=county
countrie=country
courtesie=courtesy
crazie=crazy
crie=cry
cuntrie=country
curtesie=courtesy
d'=do
daie=day
de#1=do
dear#1=deer
deare#1=deer
deer#1=dear
defie=defy
denie=deny
di'd#1=died
di'de#1=died
did'st=didst
didd=did
didd'st=didst
dide#1=died
do's=does
do'st=dost
doe=do
doe's=does
doest=dost
doeth=doth
don=done
doo=do
doo's=does
doo'st=dost
dooe=do
dooes=does
dooest=dost
dooing=doing
doost=dost
doote=do't#1
dooth=doth
dos=does
dost#1=does't#1
dowrie=dowry
drie=dry
drowsie=drowsy
dutie=duty
e'm='em
e'n#1=even
easie=easy
eie=eye
em'='em
emptie=empty
enemie=enemy
enuie=enuy
eu'n=euen
eu'ry=euery
euerie=euery
familie=family
fancie=fancy
fierie=fiery
fiftie=fifty
flie=fly
fort#1=for't#1
fortie=forty
fortifie=fortify
furie=fury
giddie=giddy
glorie=glory
goe=go
grauitie=grauity
great'st=greatest
greedie=greedy
guiltie=guilty
h'as=has
ha#1=haue
ha'=haue
ha'#1=haue
ha's=has
ha'st=hast
```

```
happie=happy
has't=hast
hast#1=haste
hastie=hasty
he'l=he'll
he'le=he'll
heard#1=herd
heau'n=heauen
heauie=heauy
hee=he
hee'd=he'd
hee'l=he'll
hee'ld=he'd
hee'le=he'll
hee'll=he'll
heel#1=he'll
heele=he'll
heer=here
heere=here
heere#1=hear
heeres=here's
hees=he's
hel#1=hell#1
here#1=hear
heres=here's
herselfe=herself
himselfe=himself
honestie=honesty
humilitie=humility
i#1=aye
i#2=in
i'=in
i'am=i'm
i'de=i'd
i'l=i'll
i'ld=i'd
i'le=i'll
i'll=i'll
i'me=i'm
i'st=is
i'th=i'th'
i'the=i'th'
iealousie=iealousy
il'd=i'd
ile=i'll
Ile#1=isle
ime=i'm
in#1=e'en
indeede=indeed
inough=enough
int=in't#1
intoo=into
ist=is't#1
it#1=its
ith=i'th'
ith'=i'th'
itselfe=itself
iustifie=iustify
ladie=lady
lazie=lazy
lecherie=lechery
lets=let's#3
libertie=liberty
liuerie=liuery
lowsie=lowsy
lowzie=lowzy
lustie=lusty
maie=may
maiestie=maiesty
manie=many
marrie=marry
mee=me
memorie=memory
mercie=mercy
merrie=merry
mightie=mighty
miserie=misery
modestie=modesty
myselfe=myself
ne're=neuer
necessitie=necessity
neu'r=neuer
no#1=not
noe=no
nobilitie=nobility
not#1=knot
nowe=now
o#1=of
o#2=he
o#6=on
o'=of
o'#1=of
```

```
o'#2=he
o'#6=on
o're=ouer
o'th=o'th'
of#1=off
off#1=of
olde=old
on#1=one
one#1=on
ons#2=on's#2
ont=on't#1
or#1=our
oth#1=oath
oth'=o'th'
our#1=ours
ourselfe=ourself
partie=party
periurie=periury
pitie=pity
pittie=pitty
plentie=plenty
policie=policy
pouertie=pouerty
praie=pray
pre-thee=prithee
pre'thee=prithee
pree-thee=prithee
preethe=prithee
preethee=prithee
prethe=prithee
prethee=prithee
prettie=pretty
priuie=priuy
prophesie=prophesy
prythee=prithee
puppie=puppy
qualitie=quality
ratifie=ratify
readie=ready
remedie=remedy
royaltie=royalty
safetie=safety
saie=say
satisfie=satisfy
sawcie=sawcy
scuruie=scuruy
selfe=self
sh'=she
shal=shall
shalbe=shallbe
she'l=she'll
she'le=she'll
shee=she
shee'd=she'd
shee'l=she'll
shee'ld=she'd
shee'le=she'll
shee'll=she'll
sheel=she'll
sheele=she'll
shees=she's
shes=she's
signifie=signify
societie=society
soe=so
sonne=son
sonnes=sons
sorrie=sorry
storie=story
studie=study
t#1=to
t#2=the
t'=to
t'is='tis
t'was='twas
t'wer='twere
t'were='twere
t'wil='twill
t'will='twill
t'would='twould
tel=tell
testifie=testify
th=the
th'=the
th'#1=thou
th'#2=they
th'ourt=thou'rt
thanck=thank
thats=that's
the#1=thee
theire=their
theis=these
```

```
ther=there
therby=thereby
there#1=their
theres=there's
therfore=therefore
thers=there's
they'l=they'll
they'ld=they'd
they'le=they'll
they'r=they're
theyl=they'll
theyle=they'll
thinck=think
thincke=think
thinke=think
thirstie=thirsty
thirtie=thirty
tho=though
thogh=though
thoud'st=thou'dst
thowlt=thou'lt
thyselfe=thyself
tis='tis
to#1=too
to#2=to
to#3=two
to#4=to
to'th=to'th'
too#1=to
too#1'th=to'th'
too#2=to
too#3=two
too#4=to
too'th=to'th'
toot#1=to't#1
toote#1=to't#1
toth'=to'th'
trie=try
twas='twas
tween=between
tweene=between
twentie=twenty
twer='twere
twere='twere
twil='twill
twilbe='twillbe
twill='twill
twold='twould
twou'd='twould
twould='twould
tydie=tydy
vanitie=vanity
verie=very
vnles=vnless
vnlesse=vnless
vntie=vnty
vntoo=vnto
vppon=vpon
was't=wast
wast#2=waste
we#1=oui
we'd=we would
we'l=we'll
we'ld=we would
we'le=we'll
wearie=weary
wee=we
wee#1=oui
wee'l=we'll
wee'le=we'll
wee'll=we'll
weed#1=we'd
weel=we'll
weele=we'll
wer=were
wer#1=wear
wer't=wert
were#1=wear
wert#1=were't#1
whats=what's
whe'r=where
whe'r#1=whether
wher=where
where#1=whether
where#2=wherever
wheres=where's
wherin=wherein
wherof=whereof
whers=where's
whie=why
whose#1=who's
wi'=will
```

```
wil=will
wilbe=willbe
wild#1=willed
wilt#1=will't#1
wo'd=would
wold=would
woulde=would
worthie=worthy
wou'd=would
y'=ye
y'#1=you
y'ar=y'are
y'ar'=y'are
y'th=i'th'
ye'=ye
yee=ye
yf=if
yle=i'll
yong=young
yor=your
you'l=you'll
you'ld=you'd
you'le=you'll
you'r=you're
youe=you'ue
youl=you'll
youl'd=you'd
youle=you'll
your#1=you're
your'#1=you're
yourselfe=yourself
yow=you
yt=it
```

Appendix D

Expansion List for Compound Contractions

This appendix contains the expansion list used with program REPLACE to expand compound contractions in the 34 plays used in this study. Many of the comments made at the beginning of Appendix C regarding the program and the format of the list apply here. For convenience, the list is here divided into two parts: the entries in first part contain an apostrophe, while those in the second usually end in a hash suffix.

The number of entries in the expansion list has been kept to a minimum by using the variant spelling list to standardize all the orthographical forms of a number of compound contractions (for example, *Ile, I'le, I'l* and *Yle* for *I'll*). In addition, each element of an expanded enclitic contraction of *is* or *it* that is expanded (according to the special use of the hash suffix system described in Section 2.5) is checked against the spelling variant list and replaced if found there. Thus, the variant spelling list given in Appendix C should always be used in conjunction with this contraction expansion list.

Expansions Containing Apostrophes:

```
'thad=it had
'thas=it has
'tis=it is
'tshallbe=it shall be
'twas=it was
'twere=it were
'twillbe=it will be
'twill=it will
'twould=it would
a'th'=on the
at'th'=at the
by'r=by your
by'th'=by the
don't=do not
h'ad=he had
ha't=haue it
he'd=he would
he'll=he will
hee'd=he would
hees=he is
i'd=i would
i'll=i will
i'm=i am
i'th'=in the
i'thy=in thy
i'ue=i have
into'th'=into the
o'both=of both
o'man=of man
o'mans=of man's#1
o'me=of me
o'mine=of mine
o'my=of my
o'that=of that
o'th'=of the
o'your=of your
oth'=of the
shallbe=shall be
she'd=she would
she'll=she will
shee'l=she will
shee'ld=she would
shees=she is
t'has=it has
t'haue=to haue
t'had=it had
th'hadst=thou hadst
th'haue=they haue
th'are=they are
th'art=thou art
they'd=they would
they'ld=they would
they'll=they will
they're=they are
theyle=they will
thou'dst=thou wouldst
thou'lt=thou wilt
thou'rt=thou art
thou'st=thou hast
thou'se=thou shalt
to'th'=to the
w'are=we are
w'haue=we haue
we'll=we will
we're=we are
wheres=where is
who'll=who will
with'=with the
willbe=will be
y'are=ye are
y'aue=ye haue
y'haue=ye haue
y'had=ye had
y'owe=ye owe
ye'are=ye are
ye'aue=ye haue
ye'haue=ye haue
ye'had=ye had
ye're=ye are
ye'ue=ye haue
yle=i will
you'd=you would
you'll=you will
you're=you are
you'ue=you haue
youle=you will
```

Expansions with Hash Suffixes:

```
'has#1=he has
at'#1=at the
ayle#1=I will
dost#1=does it
fort#1=for it
h'as#1=he has
ha's#9=he has
has#1=he has
has't#2=hast thou
hath'#1=he hath
hath#1=he hath
sha't#9=thou shalt
shal't#9=thou shalt
tone#2=the one
tother=the other
wast#1=was it
wert#1=were it
wer't#9=were it
with'#1=with the
wilt#1=will it
your#1=you are
your'#1=you are
```

Appendix E

Counts of Marker Words

This appendix is made up of tables containing word counts for every scene in the design set, the test set and the two disputed plays. Following a heading that contains the abbreviation of the play's title, each scene is identified by the act and scene number in arabic numerals. The next column is the total number of words in that scene, and the following 16 columns are the counts for the final set of 16 markers selected in Chapter 5 and analyzed in Chapter 6. All counts are from the expanded versions of the texts.

Unfortunately, there is not enough room in the tables to give complete column headings to identify each word after fitting all 16 counts into a single line. A set of one- or two-letter abbreviations is used instead. The following table provides a key to these abbreviations:

Abbr.	Marker	Abbr.	Marker	Abbr.	Marker	Abbr.	Marker
al	all	ar	are	dr	dare	dd	did
in	in	mu	must	no	no	nw	now
of	of	su	sure	th	the	ts	these
to	to	wh	which	F	Infreq-Fl+	S	Infreq-Sh+

A computer file containing these counts is freely available; please write to me or to Prof. Michaelson at one of the addresses listed in Appendix C. (Ask for file WDCTS16.) There is also a file that contains rates for these words in addition to the rates for a number of the words eliminated from consideration as markers of authorship in Chapter 5. The words include those in Table 5–3 and the *where/there* compounds, plus the rate of contraction *C* for *is*. (Ask for file WDRATES34.)

Ant		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	505	2	1	0	2	8	2	1	4	15	0	26	0	1	3	6	3
1.2	1431	5	12	1	1	20	6	7	3	26	0	49	1	2	6	16	22
1.3	860	2	4	0	3	12	1	5	5	10	0	31	0	1	5	7	11
1.4	685	4	2	0	4	7	3	3	1	14	0	27	1	2	5	7	10
1.5	610	1	0	0	4	9	1	2	4	14	0	15	0	1	1	7	5
2.1	418	2	4	0	2	5	0	1	1	8	0	16	0	0	1	7	4
2.2	1924	5	2	0	19	26	7	10	6	40	0	66	0	3	20	19	21
2.3	332	4	0	0	0	7	1	2	1	2	1	8	0	0	2	9	3
2.4	72	0	0	0	0	1	0	1	0	0	0	1	0	0	1	0	0
2.5	1014	2	4	0	1	10	0	6	2	10	1	25	1	3	2	13	12
2.6	1076	4	5	0	4	7	3	3	2	17	1	31	0	1	6	15	4
2.7	1074	8	8	0	0	12	2	4	3	17	0	52	6	1	3	14	7
3.1	290	0	0	0	0	5	1	0	3	7	0	11	0	2	3	3	4
3.2	535	2	3	0	1	6	0	3	0	12	0	24	1	1	1	6	5
3.3	390	1	5	1	0	6	1	3	0	2	0	5	0	1	0	5	2
3.4	301	3	1	0	1	0	1	1	0	5	0	8	1	0	1	2	1
3.5	176	1	0	0	0	3	0	1	1	5	0	10	0	0	0	3	1
3.6	751	3	3	0	4	13	1	1	3	27	0	32	1	1	3	13	13
3.7	650	1	6	0	0	11	1	1	1	6	0	24	3	0	4	4	12
3.8	32	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0
3.9	30	0	0	0	0	1	0	0	0	3	0	3	0	0	1	0	1
3.10	284	3	3	0	2	4	0	1	0	10	1	19	0	0	0	2	3
3.11	566	2	0	0	1	5	1	9	3	10	0	23	0	1	3	6	3
3.12	288	0	1	0	0	5	0	1	2	6	0	8	0	0	3	2	6
3.13	1649	6	3	2	4	20	2	4	4	30	2	48	1	2	2	22	23
4.1	138	0	1	0	0	1	1	1	1	4	0	4	0	0	0	1	1
4.2	384	2	0	0	1	3	0	2	1	7	0	6	0	3	1	5	4
4.3	168	0	0	0	0	1	0	1	4	2	0	6	0	0	0	2	2
4.4	316	0	0	0	0	2	1	1	3	5	0	5	0	1	0	1	1
4.5	144	0	0	0	0	0	0	1	0	1	0	3	0	0	0	4	1
4.6	282	1	0	0	2	1	1	3	1	8	0	13	0	0	1	3	1
4.7	134	0	1	0	0	0	0	0	1	0	0	0	0	1	0	6	0
4.8	325	5	1	0	0	2	0	0	0	7	0	14	0	0	1	4	4
4.9	259	1	0	0	1	3	1	1	0	7	0	12	0	1	1	2	2
4.10	72	0	0	0	0	2	0	0	0	0	0	5	0	1	0	1	1
4.11	34	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	2
4.12	410	8	1	1	0	1	0	1	0	6	0	12	0	0	0	9	4
4.13	85	0	1	0	0	1	0	0	0	1	0	6	0	0	0	1	1
4.14	1144	8	3	0	4	10	3	3	10	13	0	33	2	2	8	14	10
4.15	719	3	2	3	1	5	2	8	7	13	0	27	0	1	2	13	5
5.1	628	3	1	0	4	14	3	1	1	17	0	22	0	0	2	9	3
5.2	2932	10	8	0	3	37	7	17	10	67	2	94	4	5	15	35	20

AWW		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	1778	6	5	0	1	37	8	8	4	37	0	54	2	4	10	20	15
1.2	609	1	5	0	2	11	0	1	3	9	0	15	1	2	2	10	3
1.3	2061	6	14	1	2	36	3	13	5	39	2	61	1	0	6	29	11
2.1	1689	7	2	2	1	18	5	9	3	30	1	43	1	4	1	23	18
2.2	545	7	1	0	0	4	2	1	1	8	0	18	0	0	0	11	1
2.3	2406	10	11	3	3	34	4	7	4	48	4	72	5	10	13	33	24
2.4	431	0	1	0	2	7	0	0	1	5	0	11	0	0	3	11	5
2.5	725	1	2	1	0	6	2	2	0	11	0	12	1	0	1	12	9
3.1	179	1	0	1	0	2	0	0	1	5	1	9	0	0	0	1	4
3.2	1047	9	4	0	1	12	0	9	0	26	0	41	0	3	5	7	5
3.3	95	0	0	0	0	1	0	0	0	4	0	2	0	1	0	1	1
3.4	344	0	1	0	1	4	0	1	0	7	0	6	0	2	2	6	6
3.5	799	5	5	0	1	7	0	2	1	15	0	37	2	1	3	6	4
3.6	975	5	1	0	1	22	2	5	4	24	1	32	0	1	5	9	5
3.7	401	1	1	0	0	10	0	1	3	4	0	8	1	1	3	5	3
4.1	704	3	3	1	0	4	8	4	2	19	0	24	2	2	2	8	6
4.2	662	2	7	0	2	14	0	9	3	6	0	17	0	0	3	9	5
4.3	2667	11	13	1	0	48	4	10	6	78	0	107	2	2	5	24	26
4.4	300	1	0	0	1	2	3	0	0	5	0	14	0	0	3	6	2
4.5	855	1	4	0	1	7	0	10	0	26	1	30	0	2	5	6	5
5.1	309	1	2	0	0	4	2	1	0	3	0	9	0	0	3	3	3
5.2	433	0	1	0	1	5	0	1	3	10	0	11	0	1	0	0	5
5.3	2824	17	13	0	15	32	4	9	8	45	1	68	3	6	11	38	26
Epi	52	1	0	0	0	0	0	0	1	0	0	2	0	0	1	1	0

AYL		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	1434	3	8	0	1	17	3	9	3	27	0	35	0	1	5	12	12
1.2	2157	8	10	0	2	26	4	14	9	40	1	82	3	4	6	33	13
1.3	1092	3	6	0	5	14	0	8	3	15	0	18	3	3	2	11	8
2.1	539	0	4	0	6	12	0	1	1	16	0	28	3	1	4	7	8
2.2	160	0	2	0	2	3	0	1	0	6	0	8	1	0	0	4	1
2.3	615	4	3	0	3	7	1	6	2	12	0	20	2	2	1	9	4
2.4	790	3	4	0	3	14	2	4	6	14	1	15	1	0	0	9	5
2.5	398	3	0	0	0	2	0	2	0	6	0	11	0	1	0	4	2
2.6	171	0	0	0	0	3	0	2	1	2	0	2	0	0	0	1	1
2.7	1575	8	4	0	2	26	5	3	2	26	0	54	0	1	3	17	13
3.1	147	1	0	0	0	3	0	1	0	4	0	1	0	0	0	1	0
3.2	3206	13	19	0	4	57	8	26	7	88	2	112	6	8	12	25	18
3.3	796	1	5	0	0	9	3	11	1	16	0	28	0	0	0	6	3
3.4	450	1	2	0	1	9	0	2	0	20	0	18	0	0	0	5	1
3.5	1196	6	4	0	3	23	2	7	9	13	3	27	0	2	2	19	2
4.1	1668	9	17	0	1	25	4	8	6	41	0	45	3	2	9	14	8
4.2	126	0	0	0	0	1	0	3	0	1	0	8	0	0	1	0	0
4.3	1411	4	5	0	16	16	2	6	3	20	1	38	0	2	4	8	14
5.1	468	2	2	0	0	12	1	2	2	4	0	16	0	0	5	2	4
5.2	995	13	5	0	1	9	0	12	1	22	0	20	1	1	2	9	4
5.3	290	1	4	0	1	7	0	2	0	3	0	15	1	0	1	1	0
5.4	1552	9	6	0	3	23	3	7	2	23	2	61	8	1	2	15	10
Epi	229	0	0	0	0	2	0	3	0	5	1	13	0	0	0	3	0

CE		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	1200	4	1	0	4	14	2	4	3	34	0	49	0	1	5	21	22
1.2	874	1	3	0	0	8	0	3	4	15	0	38	1	2	0	9	12
2.1	945	1	4	0	1	11	2	7	2	10	3	16	1	2	0	11	14
2.2	1716	9	1	0	5	34	1	10	5	19	3	42	2	4	1	17	19
3.1	1198	4	6	0	1	25	3	4	1	12	0	37	0	4	1	9	11
3.2	1529	3	3	0	2	27	0	9	2	23	0	36	0	0	0	14	20
4.1	965	2	0	0	1	11	2	2	4	8	0	40	0	4	1	3	4
4.2	595	1	0	0	6	18	0	6	2	4	0	16	0	1	1	7	1
4.3	772	1	5	0	0	6	2	1	5	16	1	30	1	1	0	4	10
4.4	1318	5	5	0	13	20	1	4	6	13	0	31	6	1	0	10	14
5.1	3434	12	18	1	23	49	1	8	12	61	6	92	18	3	13	40	36

Cor		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	2200	10	18	0	7	25	8	8	2	35	0	122	7	3	10	35	24
1.2	306	0	3	0	1	6	1	1	0	5	0	9	2	0	2	7	2
1.3	914	3	4	0	3	16	3	6	2	10	0	17	0	0	0	13	14
1.4	553	3	3	0	1	4	0	3	4	10	0	20	1	0	2	15	6
1.5	218	1	0	0	0	2	0	1	1	4	0	10	2	1	0	7	2
1.6	727	4	7	1	3	13	1	0	0	16	0	28	1	2	4	8	6
1.7	64	0	0	0	0	0	0	0	0	0	0	5	0	0	0	1	1
1.8	126	0	0	0	0	2	0	0	0	1	0	6	1	0	0	1	0
1.9	762	7	2	0	1	9	2	5	2	14	0	29	1	1	4	6	5
1.10	281	1	1	0	0	4	1	0	0	5	0	11	0	0	0	9	2
2.1	2144	10	20	0	1	47	7	11	6	47	1	100	3	8	5	30	16
2.2	1312	7	4	0	5	15	1	4	4	27	0	55	0	0	5	34	14
2.3	2100	8	9	0	9	23	7	16	7	42	2	70	1	1	10	38	14
3.1	2699	13	24	1	6	29	9	20	8	49	2	129	4	5	15	48	28
3.2	1246	5	6	0	1	25	8	4	8	22	0	43	1	5	8	31	11
3.3	1132	4	2	0	0	17	2	5	1	20	0	51	0	0	4	20	9
4.1	484	2	1	0	0	4	0	0	1	10	0	16	2	1	1	9	8
4.2	470	3	3	0	0	6	0	1	2	6	0	10	0	2	1	9	5
4.3	397	2	4	0	0	9	0	2	2	8	0	18	1	0	0	11	6
4.4	219	0	1	0	0	3	0	0	1	3	0	4	1	0	1	1	2
4.5	1800	8	10	0	2	26	0	7	3	43	0	53	0	5	3	20	12
4.6	1265	11	10	1	10	12	0	4	3	18	0	50	2	0	4	13	9
4.7	481	5	3	1	1	8	1	3	1	12	1	21	1	1	3	8	3
5.1	632	2	4	0	1	6	2	2	0	10	1	13	2	1	1	8	4
5.2	888	1	10	0	0	12	3	4	4	19	0	22	0	0	0	7	11
5.3	1737	6	5	1	2	19	4	10	4	29	1	66	2	1	11	21	16
5.4	533	3	3	0	2	7	0	5	1	12	0	26	0	0	0	7	2
5.5	44	1	0	0	0	0	0	0	0	2	0	4	0	0	0	1	0
5.6	1247	5	3	0	6	14	3	8	1	26	0	47	0	1	8	13	9

Cym		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	1495	4	1	1	4	15	3	7	2	29	0	58	1	3	4	18	24
1.2	272	0	0	0	0	2	0	3	0	5	0	7	0	0	1	2	0
1.3	340	2	0	0	3	2	0	1	0	8	0	13	0	0	1	3	4
1.4	1396	3	8	2	1	24	2	8	2	37	0	40	2	8	8	8	17
1.5	735	2	2	0	1	4	0	4	3	14	0	20	3	1	5	14	7
1.6	1671	9	10	0	2	26	4	4	0	26	2	54	0	2	11	18	18
2.1	489	2	2	1	1	2	2	4	0	13	0	9	0	1	0	6	3
2.2	418	1	1	0	1	3	0	1	1	13	0	32	1	0	0	8	2
2.3	1282	4	7	0	0	8	4	9	4	28	0	38	0	6	7	19	10
2.4	1298	7	6	0	2	12	7	8	4	31	2	43	1	4	6	21	12
2.5	287	5	3	0	2	5	1	3	0	5	0	8	0	0	1	8	3
3.1	687	2	1	0	4	11	1	4	3	10	0	20	0	1	9	16	8
3.2	697	3	3	0	0	14	0	3	2	16	0	13	1	3	1	9	6
3.3	924	2	7	0	2	17	1	2	1	23	0	47	4	0	3	19	9
3.4	1618	11	6	0	1	20	8	11	5	35	0	52	0	2	4	17	15
3.5	1362	7	4	1	0	16	5	4	3	30	0	39	2	1	5	12	12
3.6	810	1	2	1	1	10	0	6	2	7	0	20	1	0	1	8	5
3.7	117	0	2	0	0	3	1	0	3	5	0	12	0	0	1	0	0
4.1	241	2	1	1	0	7	1	1	1	5	0	10	0	1	1	5	1
4.2	3392	15	20	0	7	29	7	20	8	50	2	120	6	1	11	65	34
4.3	405	4	3	1	1	7	2	4	1	10	0	13	1	0	1	8	7
4.4	462	0	2	0	1	6	1	2	1	10	0	18	0	1	1	10	6
5.1	283	1	0	0	0	2	1	3	0	5	0	10	1	0	0	5	6
5.2	148	0	3	0	0	1	0	0	0	5	0	10	0	0	0	1	1
5.3	820	3	6	0	4	14	1	3	4	11	0	46	1	2	5	11	7
5.4	1531	5	10	1	1	17	2	10	2	32	3	43	2	6	5	27	20
5.5	4008	21	16	0	19	44	6	11	12	91	3	130	12	8	32	41	51

1H4		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	812	4	2	0	7	14	1	4	4	36	0	30	1	0	4	10	8
1.2	1794	5	3	0	2	24	1	10	9	45	0	78	0	3	3	26	12
1.3	2316	10	3	0	14	30	0	10	3	53	1	76	6	3	11	33	21
2.1	784	4	7	0	0	14	0	6	0	19	0	28	1	1	1	5	6
2.2	845	6	6	1	0	7	1	4	4	7	0	28	1	1	0	17	6
2.3	1002	4	7	0	0	14	8	4	4	26	0	24	2	3	0	10	6
2.4	4395	23	16	0	6	71	3	13	21	103	1	132	9	6	3	52	28
3.1	2062	9	6	0	2	44	3	10	3	47	2	78	7	5	5	18	32
3.2	1374	11	4	0	3	22	2	4	4	45	0	43	1	1	10	16	18
3.3	1734	6	3	1	3	31	4	9	7	42	1	41	2	1	0	23	8
4.1	1094	12	1	1	2	16	2	7	5	38	0	45	1	2	0	12	16
4.2	676	6	3	0	1	7	1	3	4	16	1	26	1	3	0	6	3
4.3	881	4	5	1	4	18	0	1	3	26	0	37	1	3	3	6	7
4.4	314	2	1	0	0	3	2	2	0	10	0	16	0	2	0	4	2
5.1	1139	6	2	1	8	20	1	13	4	32	0	39	1	1	1	12	9
5.2	780	8	3	1	9	10	1	4	4	27	0	28	0	1	3	14	6
5.3	517	2	4	0	1	7	0	4	2	7	1	12	0	2	1	8	1
5.4	1417	6	3	0	7	19	1	7	4	36	2	43	1	7	2	13	4
5.5	332	4	1	0	2	3	0	0	0	10	0	16	0	1	1	6	3

H5		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.P	245	1	1	0	1	3	1	0	2	7	0	13	1	0	1	2	0
1.1	737	5	3	0	4	14	4	3	3	23	0	39	0	1	6	11	7
1.2	2428	13	9	0	10	61	2	8	5	77	1	124	4	1	9	18	33
2.P	322	3	2	0	0	6	2	0	6	11	0	22	0	0	1	0	3
2.1	980	1	2	1	1	11	7	2	4	18	0	34	1	0	1	11	4
2.2	1488	6	6	1	5	20	2	6	5	60	0	41	2	3	6	16	16
2.3	474	2	2	0	4	8	1	2	2	9	1	13	0	0	0	4	2
2.4	1138	6	3	0	1	24	0	7	2	44	0	46	1	3	1	16	16
3.P	261	1	0	0	0	2	0	1	1	3	0	18	1	0	1	5	3
3.1	275	0	1	0	1	6	0	0	3	8	0	20	1	0	1	4	4
3.2	1181	5	4	0	0	18	3	3	2	30	0	68	2	2	1	16	10
3.3	449	3	5	0	1	9	0	1	0	16	0	25	0	0	0	9	1
3.5	522	2	2	0	0	11	0	0	2	14	1	14	0	0	0	4	5
3.6	1422	3	5	0	3	16	5	3	8	33	0	57	0	2	5	12	5
3.7	1176	3	5	1	3	10	1	3	2	39	0	50	1	0	2	14	8
4.P	383	2	0	0	0	4	1	1	2	12	0	27	0	0	0	6	11
4.1	2495	17	13	1	0	34	3	14	7	71	1	122	4	3	3	33	18
4.2	469	5	2	1	0	10	1	0	1	8	0	19	0	0	0	13	6
4.3	1094	8	8	0	2	21	2	7	7	23	0	37	2	0	4	15	10
4.4	372	0	2	0	1	4	1	0	0	9	0	12	0	0	0	0	1
4.5	166	3	2	0	0	5	0	1	3	0	0	7	1	1	0	4	2
4.6	299	4	0	0	2	6	1	0	0	5	0	12	1	1	2	7	3
4.7	1509	7	2	1	4	32	1	4	5	38	0	62	1	1	3	16	13
4.8	999	4	4	0	2	20	1	2	8	39	0	37	2	0	3	2	10
5.P	354	2	0	0	1	8	1	0	5	10	0	22	0	0	2	7	7
5.1	731	5	1	1	0	12	1	4	2	14	0	9	1	1	2	8	5
5.2	3053	16	14	1	1	51	8	12	6	74	1	86	3	4	14	40	19

JC		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	588	4	2	0	0	11	1	3	3	10	0	18	2	1	1	9	4
1.2	2588	7	9	0	15	31	2	10	9	48	2	78	4	6	5	37	27
1.3	1299	10	12	0	6	19	1	5	5	25	1	49	10	1	2	23	9
2.1	2634	15	14	1	8	24	6	23	3	64	2	92	6	5	11	38	28
2.2	1031	4	7	1	5	18	0	2	5	11	0	33	3	1	4	19	10
2.3	117	1	0	0	0	1	0	0	0	4	0	3	1	0	0	0	0
2.4	385	0	0	0	0	2	1	0	0	3	1	13	0	0	1	5	4
3.1	2307	17	11	0	6	32	5	12	9	58	0	67	5	1	4	27	21
3.2	2124	14	13	0	5	20	3	7	6	34	1	68	1	0	7	31	9
3.3	250	1	3	0	1	0	0	2	0	2	0	4	0	0	0	1	2
4.1	395	0	3	0	0	6	3	0	1	6	0	9	2	1	1	4	2
4.2	365	0	2	0	0	6	0	2	1	5	0	8	0	0	1	8	3
4.3	2378	7	16	0	13	24	11	14	7	37	2	43	2	5	7	29	23
5.1	971	3	7	1	3	11	2	3	5	19	0	36	0	3	3	21	5
5.2	45	1	0	0	0	1	0	0	0	0	0	3	1	0	0	3	1
5.3	866	3	9	0	7	8	0	3	6	11	0	21	2	3	0	20	7
5.4	248	1	0	1	0	2	1	2	1	2	0	6	0	0	0	5	4
5.5	641	8	1	0	5	9	0	5	2	12	1	16	0	0	0	6	7

KJ		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	2213	8	1	0	6	22	7	9	11	47	0	62	3	2	8	16	21
2.1	4679	21	16	0	5	93	2	10	12	155	0	155	18	2	12	50	61
3.1	2766	9	5	0	2	27	5	13	5	63	0	79	6	2	11	34	21
3.2	74	0	0	0	0	3	0	0	1	0	0	1	0	0	0	1	1
3.3	598	2	0	0	1	5	1	1	2	10	0	11	0	2	1	5	1
3.4	1474	14	5	0	3	22	3	12	7	43	0	36	4	3	5	16	14
4.1	1145	5	6	0	4	13	9	9	2	16	0	24	7	1	3	16	8
4.2	2132	10	5	0	5	28	5	6	2	63	1	84	3	0	8	31	22
4.3	1299	5	1	1	1	13	2	2	6	38	0	54	2	1	1	18	13
5.1	605	2	3	0	4	5	1	1	1	17	0	32	0	0	0	10	4
5.2	1406	3	1	0	0	17	3	5	6	43	0		6	5	2	19	18
5.3	136	0	1	0	0	0	0	0	1	1	0		0	0	1	1	0
5.4	483	3	1	0	1	8	2	0	1	17	0	40	0	1	2	6	7
5.5	184	0	2	0	3	2	0	0	0	4	0	9	0	0	1	6	0
5.6	358	3	3	0	2	2	0	1	0	9	0	15	2	0	0	5	1
5.7	951	6	4	0	3	12	2	0	11	22	0	39	2	1	6	17	13

LLL		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	2325	11	12	0	5	39	3	11	4	55	0	97	7	7	16	17	27
1.2	1266	1	6	0	0	18	3	7	2	27	2	39	0	5	9	12	6
2.1	1975	15	10	0	9	34	2	8	8	46	0	53	1	9	10	16	17
3.1	1511	6	4	0	1	22	8	15	4	32	0	55	6	2	2	17	12
4.1	1322	1	2	0	5	11	2	5	4	19	0	63	2	2	6	11	7
4.2	1401	3	11	0	2	22	0	2	2	39	0	87	1	2	5	11	11
4.3	2957	20	17	1	11	53	2	14	9	54	1	102	9	7	3	28	27
5.1	1091	3	2	0	1	6	1	2	3	28	0	63	0	6	3	8	5
5.2	7172	28	38	1	19	121	5	29	13	135	1	258	18	16	12	78	43

Mac		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	63	0	0	0	0	2	0	0	0	1	0	6	0	0	0	2	0
1.2	486	1	0	0	0	2	1	2	1	11	0	24	1	1	2	3	4
1.3	1154	7	12	0	3	16	0	2	0	33	0	46	2	1	6	18	9
1.4	472	2	3	0	1	9	3	3	1	9	0	18	0	0	5	8	5
1.5	604	3	0	0	0	9	2	1	1	17	0	31	1	1	6	11	6
1.6	254	1	1	0	0	2	1	1	0	4	0	8	0	0	1	8	1
1.7	696	2	0	2	2	13	1	2	4	12	0	36	1	0	4	12	12
2.1	529	3	2	0	0	8	0	1	3	8	0	27	0	2	7	14	6
2.2	627	2	5	1	6	5	3	6	1	10	0	27	4	0	2	16	9
2.3	1190	4	4	0	6	26	0	4	1	23	0	59	0	3	3	20	10
2.4	333	0	1	0	2	2	0	1	1	10	0	16	0	0	2	6	3
3.1	1159	4	8	0	2	25	5	4	5	26	1	45	0	0	8	24	14
3.2	450	2	2	0	0	8	2	0	1	10	0	16	2	0	3	13	4
3.3	184	1	2	0	2	1	0	0	1	4	0	12	0	0	0	2	3
3.4	1207	9	5	4	4	13	2	4	7	15	0	55	1	3	8	13	11
3.6	414	2	0	0	3	3	1	0	1	10	0	18	1	3	2	3	6
4.1	1010	2	4	0	1	11	0	7	3	35	1	46	1	1	2	13	7
4.2	676	6	6	2	0	5	4	4	3	5	0	19	1	2	1	17	7
4.3	1930	15	7	1	6	21	3	8	5	35	1	68	4	6	5	31	21
5.1	598	5	3	1	0	9	0	4	3	11	1	16	1	0	2	12	7
5.2	247	2	1	0	0	4	0	0	4	5	0	12	0	0	0	2	6
5.3	493	3	1	1	0	1	2	2	1	14	0	21	1	0	4	10	8
5.4	169	0	2	0	0	2	1	1	0	3	0	8	0	1	1	2	1
5.5	426	1	0	0	1	3	0	1	2	8	0	19	0	0	2	8	3
5.6	81	2	1	0	0	0	0	0	1	1	0	1	0	0	0	1	0
5.7	246	0	1	0	0	2	1	3	0	5	0	10	0	0	0	4	0
5.8	638	2	1	0	0	6	3	6	0	14	0	23	2	1	3	9	6

MAN		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	2487	12	9	1	1	46	3	14	6	45	0	75	2	8	1	29	17
1.2	216	0	1	0	0	3	0	2	1	4	0	7	0	0	0	3	0
1.3	564	3	2	0	0	9	3	6	0	10	1	18	0	0	1	3	5
2.1	2953	10	12	0	5	42	2	19	7	46	4	99	1	14	6	23	16
2.2	416	1	0	0	0	11	0	2	0	11	0	20	0	0	1	8	1
2.3	2120	9	4	0	4	29	5	19	7	41	1	50	3	1	1	22	19
3.1	927	4	3	1	5	10	2	5	3	18	4	22	0	1	3	10	7
3.2	933	3	2	2	1	13	2	8	3	15	0	27	1	3	3	4	4
3.3	1342	6	11	0	3	14	1	7	1	23	0	65	1	2	4	18	6
3.4	704	3	4	0	0	11	0	6	3	13	0	17	1	0	0	6	4
3.5	484	3	6	0	0	6	3	2	2	6	0	9	1	1	0	8	0
4.1	2545	11	10	3	4	36	3	12	5	55	1	48	6	5	5	33	13
4.2	613	1	11	0	1	7	1	1	0	4	0	17	2	0	5	8	2
5.1	2626	10	8	6	8	30	3	15	6	34	0	46	4	5	13	36	12
5.2	767	2	2	0	1	22	4	6	3	15	0	19	1	3	6	8	3
5.3	207	1	0	0	0	2	0	0	3	6	0	10	0	0	2	2	4
5.4	1055	5	6	0	6	13	3	9	0	15	1	21	0	0	5	4	8

MND		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	1933	5	2	0	6	29	5	6	3	37	0	53	1	2	8	20	17
1.2	777	6	2	0	0	13	5	5	2	11	0	35	0	3	1	6	2
2.1	2114	8	8	0	1	40	3	4	5	37	0	112	2	2	4	14	22
2.2	1158	6	6	0	3	10	3	13	7	18	0	27	0	0	0	13	10
3.1	1507	6	4	0	1	12	13	6	3	23	0	45	0	2	1	21	16
3.2	3638	22	14	0	11	50	8	20	16	41	1	67	7	6	4	48	42
4.1	1704	7	9	0	7	29	2	6	13	43	1	70	9	1	3	20	12
4.2	356	3	3	0	0	4	1	3	0	5	0	12	2	1	0	4	2
5.1	3032	23	20	0	8	53	4	22	18	48	1	127	10	2	17	39	30

MV		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	1496	8	12	0	2	25	4	3	8	34	1	39	2	3	5	20	17
1.2	1146	1	4	0	0	18	0	4	1	29	0	50	5	1	1	15	9
1.3	1419	3	4	0	5	21	0	7	1	29	0	48	4	1	5	16	7
2.1	365	1	0	0	0	2	1	0	0	9	0	22	0	1	2	2	4
2.2	1685	2	6	0	1	19	4	7	4	34	3	57	2	4	3	9	10
2.3	163	0	0	0	0	2	0	0	0	2	0	1	1	0	0	0	0
2.4	291	2	0	1	0	6	1	0	1	3	0	7	0	0	0	2	3
2.5	441	1	2	0	1	6	0	1	0	6	0	12	0	0	0	6	3
2.6	535	3	9	0	1	3	2	3	2	5	0	17	0	2	1	5	5
2.7	628	8	1	0	0	16	3	2	2	13	0	23	2	2	1	7	2
2.8	412	1	0	0	1	8	0	0	0	8	1	19	0	1	1	4	3
2.9	771	1	3	0	2	7	2	4	1	14	0	28	2	2	3	5	9
3.1	1046	2	1	1	1	14	0	7	5	25	1	42	1	1	0	9	4
3.2	2584	16	12	0	8	43	1	9	6	49	1	98	4	4	7	26	32
3.3	301	1	0	0	0	1	0	3	0	6	1	12	1	0	0	2	1
3.4	667	7	2	0	1	12	3	2	3	23	0	18	2	1	7	10	5
3.5	732	3	9	0	0	16	1	3	2	17	0	32	0	2	0	9	4
4.1	3718	16	19	0	2	44	11	21	5	86	0	176	1	2	14	38	44
4.2	161	0	1	0	2	0	0	0	0	0	0	2	0	1	1	3	1
5.1	2537	7	16	2	21	37	2	15	6	51	2	111	4	3	9	25	32

MWW		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	2220	8	7	0	6	29	4	8	7	35	0	58	6	2	3	14	15
1.2	103	0	0	0	0	1	0	0	0	3	0	3	0	0	2	1	0
1.3	756	3	0	0	3	6	4	4	4	21	0	30	2	3	2	4	2
1.4	1272	6	3	0	0	25	1	9	2	17	0	25	0	2	0	11	11
2.1	1779	2	7	0	2	23	0	5	8	37	4	53	5	2	0	18	11
2.2	2506	16	8	0	0	25	3	8	4	48	1	55	0	6	7	23	24
2.3	666	1	2	0	0	3	1	1	3	11	0	12	0	0	1	0	2
3.1	816	2	4	0	1	7	0	4	3	16	0	21	0	1	1	7	6
3.2	698	2	3	0	0	4	2	3	3	11	2	11	1	2	0	2	6
3.3	1718	4	7	0	0	38	3	2	9	29	0	51	3	6	2	16	11
3.4	857	2	0	0	0	10	4	5	7	10	0	6	1	1	0	10	10
3.5	1175	1	0	0	3	26	1	3	3	27	0	31	1	0	0	8	9
4.1	470	0	3	0	0	4	1	4	2	11	2	8	0	1	1	3	1
4.2	1602	6	8	0	2	19	2	13	6	31	3	60	0	4	2	30	9
4.3	98	0	1	0	0	1	1	0	0	2	0	4	0	0	0	4	1
4.4	739	6	1	0	3	17	1	3	1	18	1	22	1	1	0	10	6
4.5	944	3	6	0	0	6	0	4	3	26	1	30	1	2	0	9	11
4.6	445	3	2	0	0	7	2	0	2	5	0	20	0	0	2	2	8
5.1	266	1	0	0	0	9	0	1	1	5	0	6	0	0	0	2	2
5.2	117	0	0	0	0	2	0	1	0	1	0	5	0	1	0	0	0
5.3	185	1	1	0	0	2	1	2	1	4	0	11	0	0	1	1	0
5.4	33	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
5.5	1920	7	8	0	3	26	3	6	14	41	1	63	3	3	2	16	8

R2		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	1617	5	3	0	3	24	3	6	3	39	0	51	3	3	11	15	18
1.2	578	5	1	0	0	13	2	2	0	9	0	18	0	0	3	5	4
1.3	2407	7	4	0	2	40	4	12	5	66	0	77	3	6	8	24	28
1.4	481	1	3	0	4	7	1	0	2	11	0	12	2	2	2	4	1
2.1	2386	11	10	1	4	39	7	14	13	54	0	69	4	6	6	20	26
2.2	1170	8	7	0	1	9	1	4	4	14	0	35	2	3	7	18	9
2.3	1381	9	9	0	2	21	4	5	5	37	1	38	3	1	8	15	8
2.4	193	3	3	0	0	4	0	1	0	4	0	14	1	0	0	2	3
3.1	344	1	0	0	1	6	1	2	0	11	0	11	1	1	0	4	6
3.2	1755	14	6	0	1	20	4	4	3	37	0	60	1	6	5	20	18
3.3	1649	6	7	1	0	8	7	6	3	45	0	69	0	2	2	21	21
3.4	866	4	5	0	2	16	1	5	1	36	0	27	2	2	4	10	9
4.1	2642	20	14	2	6	35	4	18	7	68	0	78	6	1	3	39	25
5.1	845	2	1	0	1	12	6	1	1	13	0	33	0	2	4	13	12
5.2	967	2	4	0	2	5	3	4	5	13	0	29	2	0	4	14	12
5.3	1172	5	3	0	2	14	0	5	3	11	0	36	1	0	3	18	10
5.4	95	0	0	0	2	0	0	2	0	1	0	4	1	0	0	0	1
5.5	985	0	4	1	2	16	2	5	5	22	0	34	3	1	2	16	8
5.6	412	3	0	0	2	5	0	0	0	15	0	14	0	0	0	6	4

R3		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	1261	3	10	0	0	17	6	7	5	19	0	40	4	1	4	13	4
1.2	2064	10	8	0	6	21	0	9	4	39	0	50	12	2	8	25	15
1.3	2867	17	13	2	7	35	4	10	11	62	0	67	2	7	8	45	28
1.4	2167	2	8	0	4	35	2	13	8	48	0	73	2	1	3	17	17
2.1	1123	8	1	0	8	16	1	2	3	29	0	26	1	2	1	13	11
2.2	1210	6	4	0	3	16	1	3	2	22	0	33	2	0	3	7	17
2.3	400	6	2	0	0	4	0	5	1	8	0	17	0	2	1	1	3
2.4	598	5	3	0	4	4	0	4	2	6	0	18	0	1	0	4	4
3.1	1590	12	3	0	3	18	2	9	5	30	0	48	0	5	5	17	14
3.2	1001	2	8	0	1	8	1	4	5	10	1	43	2	1	2	7	7
3.3	197	1	1	0	0	1	1	0	2	5	0	6	0	0	1	2	1
3.4	868	4	2	0	4	16	0	4	5	21	0	31	1	3	3	12	4
3.5	819	5	5	0	1	11	1	1	3	21	0	42	1	1	6	16	3
3.6	118	1	0	0	0	3	1	0	0	1	0	6	1	0	1	2	1
3.7	1880	7	4	0	4	33	1	13	5	51	1	66	1	1	8	27	26
4.1	851	4	0	0	1	7	2	7	2	23	0	31	0	0	3	9	14
4.2	799	3	3	0	2	11	1	3	5	6	0	19	1	0	2	13	5
4.3	443	1	1	0	2	8	1	0	1	10	0	20	0	0	1	14	4
4.4	4270	20	11	0	4	54	4	13	16	104	0	135	4	9	8	51	37
4.5	158	0	0	0	0	3	0	0	1	5	0	5	0	0	0	3	2
5.1	225	1	0	0	0	4	0	1	0	5	0	8	0	0	2	2	2
5.2	189	1	1	0	0	6	0	1	1	7	0	7	0	0	1	1	3
5.3	2760	12	4	0	2	58	3	8	2	70	0	111	8	2	4	28	28
5.4	111	1	0	0	0	2	0	0	0	4	0	6	0	0	0	0	0
5.5	316	2	2	0	0	4	0	0	3	8	0	15	1	0	0	4	2

Rom		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.P	106	0	0	0	0	2	0	0	1	5	0	5	1	0	2	0	1
1.1	1828	7	1	1	2	28	2	9	1	44	0	56	3	5	5	20	29
1.2	796	5	6	0	0	11	1	1	3	15	0	25	1	1	1	9	4
1.3	894	6	3	0	1	14	2	4	7	17	0	30	1	3	0	10	6
1.4	903	0	4	0	1	16	1	4	1	33	0	35	0	7	5	11	9
1.5	1175	6	6	0	2	10	3	1	9	17	0	32	1	9	3	7	7
2.P	114	0	0	0	0	2	1	0	3	1	0	1	0	0	1	2	3
2.1	347	0	0	0	0	5	1	0	1	3	0	7	1	2	0	2	0
2.2	1573	6	3	0	3	12	1	6	0	20	0	34	1	8	7	28	12
2.3	759	4	0	0	2	13	2	3	2	9	0	18	2	0	0	13	20
2.4	1630	0	1	1	2	19	2	6	9	27	3	56	4	2	3	5	3
2.5	681	2	2	0	2	7	2	3	3	8	0	19	0	0	2	8	5
2.6	291	0	1	1	0	9	0	0	0	5	0	11	1	3	1	2	1
3.1	1591	7	2	0	3	15	4	6	4	30	0	51	2	1	5	18	15
3.2	1148	11	5	0	5	15	0	8	1	12	0	19	4	0	2	12	1
3.3	1393	6	2	0	1	20	2	7	3	22	0	31	0	1	5	19	24
3.4	301	1	0	0	1	1	0	3	0	3	0	0	1	1	0	2	2
3.5	2015	5	9	0	1	26	4	9	11	17	0	48	3	6	4	32	20
4.1	1041	2	2	0	0	16	4	7	5	20	1	26	0	2	5	7	12
4.2	373	3	2	0	0	1	0	1	3	3	0	6	0	0	0	6	2
4.3	472	4	4	0	1	9	1	4	1	7	1	14	1	0	2	10	1
4.4	231	1	1	0	0	2	0	1	3	0	0	8	1	0	0	2	1
4.5	1092	8	4	0	1	14	2	6	5	2	0	25	2	1	0	20	7
5.1	697	2	2	0	3	14	2	4	2	11	0	17	2	0	1	5	10
5.2	225	0	0	0	1	3	1	1	1	7	0	9	1	0	0	5	3
5.3	2498	9	6	2	9	31	3	7	2	44	0	68	8	5	8	23	33

Tem		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	501	3	4	0	0	2	1	2	2	13	0	21	2	0	0	10	4
1.2	4141	29	13	0	18	53	6	29	21	84	3	159	2	3	26	53	52
2.1	2441	14	12	0	4	28	2	18	10	59	2	72	3	7	9	30	26
2.2	1559	5	3	0	2	17	1	8	10	25	0	47	3	2	4	19	10
3.1	827	3	7	1	2	8	3	5	3	14	0	17	3	2	2	11	6
3.2	1161	2	3	1	3	11	0	4	2	14	0	27	0	2	1	21	7
3.3	922	3	16	0	4	14	1	5	8	23	0	25	4	2	4	18	11
4.1	2018	14	9	0	2	15	3	7	7	31	0	52	6	3	9	29	11
5.1	2603	16	8	0	6	28	6	9	13	49	0	79	10	0	14	36	24
Epi	129	2	1	0	0	1	2	0	3	2	0	2	0	0	3	0	2
TGV		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	1147	5	3	0	2	18	2	8	3	9	0	37	0	2	1	10	13
1.2	1102	6	3	1	1	10	0	4	7	17	0	28	1	6	3	22	8
1.3	689	3	2	0	1	11	1	5	3	18	0	20	1	0	2	5	9
2.1	1253	4	12	0	1	12	0	7	5	12	0	16	4	3	2	24	13
2.2	159	0	1	0	0	1	1	1	1	1	0	5	0	0	0	0	0
2.3	549	3	0	0	1	9	0	7	8	2	0	31	0	0	0	5	3
2.4	1578	10	9	0	3	15	5	10	8	29	1	27	1	4	1	18	22
2.5	379	1	2	0	1	4	0	4	0	2	0	7	0	1	0	3	2
2.6	334	1	0	0	1	2	1	0	3	1	0	1	0	0	1	6	7
2.7	713	4	6	0	1	8	2	2	2	15	0	17	1	0	0	3	8
3.1	2899	5	7	0	1	31	4	20	13	46	1	61	0	2	13	33	52
3.2	725	0	2	1	1	9	5	0	3	8	0	13	0	0	2	2	6
4.1	553	4	8	0	0	4	0	2	1	12	0	13	3	0	3	4	5
4.2	1007	3	3	0	2	10	3	0	6	6	2	21	1	3	1	7	14
4.3	346	1	2	0	1	2	0	1	0	3	0	5	0	0	2	4	6
4.4	1699	6	3	0	7	14	1	4	4	17	1	38	1	1	7	19	19
5.1	95	0	1	0	0	0	0	0	1	0	1	6	0	0	0	1	0
5.2	411	0	5	0	1	3	0	2	2	9	1	7	1	2	1	7	5
5.3	110	0	0	0	0	0	2	0	0	1	0	4	0	0	0	0	1
5.4	1366	9	5	3	2	12	1	4	6	18	1	25	3	1	3	20	10
TN		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	323	3	1	0	1	3	0	1	2	11	0	10	1	0	1	1	3
1.2	487	0	0	0	2	10	0	2	1	10	0	13	0	0	0	1	4
1.3	1039	2	6	0	2	27	2	4	3	21	1	24	6	2	0	10	3
1.4	329	4	2	0	0	6	0	3	0	5	1	6	1	0	0	3	1
1.5	2466	2	16	0	2	37	2	23	7	53	2	69	3	4	1	26	16
2.1	395	1	1	0	0	5	1	3	0	14	0	11	0	0	1	9	5
2.2	346	0	1	0	1	5	1	1	3	8	1	8	0	1	0	1	1
2.3	1439	2	7	1	3	18	2	12	5	39	0	33	0	6	0	14	8
2.4	958	7	9	0	2	15	2	8	4	20	0	31	1	3	0	6	4
2.5	1513	4	2	0	4	18	6	6	8	38	0	39	3	2	2	17	13
3.1	1304	3	12	0	3	5	3	11	3	20	0	40	1	0	2	15	9
3.2	687	1	1	0	2	17	2	7	2	21	0	28	0	0	0	7	5
3.3	397	1	1	0	3	6	0	2	0	10	0	13	3	1	3	4	7
3.4	3103	4	7	1	4	34	4	22	15	74	3	85	0	3	5	36	13
4.1	519	1	2	0	0	6	1	2	4	7	0	6	1	0	0	12	3
4.2	953	2	4	0	1	19	0	5	1	17	0	27	0	0	1	10	3
4.3	302	2	1	0	1	1	0	1	2	4	0	9	0	1	0	8	3
5.1	3162	10	13	0	14	42	3	11	11	55	0	90	4	3	8	25	31

TS		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	2016	9	3	0	1	34	1	9	5	27	0	42	2	4	0	23	11
1.2	2276	14	7	0	1	33	3	12	6	18	1	43	2	2	1	13	20
2.1	3345	26	21	0	7	44	10	14	18	42	2	61	5	6	4	35	24
3.1	702	4	1	0	0	13	4	3	4	9	1	18	1	1	1	14	4
3.2	1973	13	4	1	1	18	5	9	7	28	0	60	3	3	5	29	17
4.1	1571	14	11	0	2	14	3	9	10	19	0	41	4	0	2	28	12
4.2	959	5	7	0	0	18	1	1	3	17	0	21	1	0	1	9	7
4.3	1593	6	5	2	7	17	0	12	2	18	1	53	4	3	1	16	8
4.4	803	3	3	0	0	10	0	3	2	10	0	21	0	1	0	10	3
4.5	638	0	0	0	0	2	0	0	5	9	0	23	0	1	2	3	1
5.1	1051	7	6	2	2	12	1	2	4	11	0	34	1	1	0	7	8
5.2	1518	4	12	0	1	7	1	6	11	11	0	20	2	3	4	11	13
i.1	1079	5	4	0	0	13	1	4	2	7	1	31	0	0	2	16	7
i.2	1168	5	5	0	1	12	1	9	4	19	0	28	5	1	3	20	9

WT		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	356	0	1	0	0	6	0	2	2	9	0	11	0	1	2	3	2
1.2	3880	15	21	6	9	44	12	19	11	55	2	96	3	9	21	49	34
2.1	1704	8	7	1	3	21	1	11	5	22	0	45	4	3	10	20	18
2.2	527	0	0	1	0	4	2	4	3	12	0	25	2	1	1	4	4
2.3	1753	6	6	0	3	14	2	7	4	34	1	53	3	2	7	25	23
3.1	173	2	0	0	0	1	0	0	0	5	0	24	1	0	0	3	0
3.2	1962	15	4	0	4	22	3	12	6	50	0	69	3	8	17	23	21
3.3	1177	1	10	0	4	10	0	2	8	18	1	56	3	2	3	22	3
4.1	267	1	0	0	0	6	0	0	8	10	0	9	0	0	2	4	2
4.2	480	0	5	0	0	2	2	3	2	14	0	17	0	1	4	4	8
4.3	995	3	5	0	0	10	3	8	3	28	0	42	3	2	2	8	4
4.4	6958	27	41	2	3	68	23	36	25	139	3	230	23	19	25	94	69
5.1	1915	12	8	0	6	17	1	10	13	27	1	54	3	5	9	37	21
5.2	1513	10	8	0	2	23	2	7	9	58	1	83	3	0	12	15	8
5.3	1310	4	1	0	4	9	0	7	5	15	0	30	2	2	6	26	16

Bond		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	1591	6	9	1	2	17	5	10	1	33	1	58	18	5	1	25	7
1.2	2421	21	18	3	1	37	9	22	8	35	1	66	9	5	4	42	11
2.1	1095	11	8	3	3	9	14	10	2	13	0	26	1	3	3	18	2
2.2	988	9	7	6	1	2	5	10	7	10	2	23	0	2	0	22	4
2.3	1223	8	14	3	0	7	3	6	9	14	0	26	9	5	3	28	2
2.4	830	13	5	3	0	3	5	9	6	12	0	18	1	3	1	27	1
3.1	715	3	2	0	0	8	1	8	7	15	0	19	4	0	0	13	11
3.2	778	11	6	1	0	5	1	3	2	10	1	28	1	2	0	16	0
3.3	236	0	0	0	0	2	2	2	1	2	1	10	0	0	1	5	2
3.4	129	1	1	0	0	0	0	1	1	1	0	6	0	1	0	1	0
3.5	1621	11	15	0	1	15	3	7	12	7	2	49	6	5	2	50	8
4.1	618	8	6	1	1	6	0	3	2	10	0	17	0	3	1	18	2
4.2	840	3	2	1	0	7	3	5	1	6	2	20	2	2	0	13	6
4.3	1896	18	10	6	0	16	6	15	10	27	3	55	2	2	2	46	11
4.4	1381	15	5	3	0	16	5	8	2	15	2	37	4	0	1	20	6
5.1	800	8	3	0	1	6	0	1	7	15	0	17	4	3	3	20	2
5.2	1425	7	6	2	2	8	4	6	4	21	1	40	2	8	0	23	6
5.3	1862	18	7	5	0	12	5	11	9	24	5	50	4	5	1	37	3

Chan		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	653	3	9	0	0	8	1	7	5	9	1	11	1	4	1	13	8
1.2	473	5	2	2	0	3	4	5	1	7	3	10	0	3	0	11	3
1.3	364	5	4	0	0	5	2	3	2	8	1	3	0	1	1	6	1
1.4	80	1	3	0	0	0	0	1	0	0	1	2	0	0	0	0	1
1.5	423	5	0	0	1	5	3	3	3	9	1	11	1	1	0	4	4
1.6	256	1	3	0	0	2	3	1	0	3	1	3	0	1	1	5	3
1.7	272	1	6	0	0	0	0	5	1	4	1	5	1	1	0	4	0
1.8	894	7	7	0	1	5	4	10	5	14	2	16	5	3	0	13	4
1.9	352	2	1	0	0	5	0	6	1	10	1	9	0	2	0	6	2
1.10	521	11	6	1	0	1	1	3	4	5	0	13	0	3	0	15	4
2.1	1237	7	6	2	1	10	4	13	5	20	7	12	2	4	4	14	3
2.2	430	1	2	0	0	5	1	5	1	1	1	8	0	2	0	9	2
2.3	1134	16	6	0	1	9	2	5	8	13	0	21	2	5	2	20	4
2.4	777	13	2	2	0	9	3	4	2	14	1	11	2	2	4	12	10
3.1	1165	8	6	3	2	5	6	10	8	6	1	14	3	8	1	23	2
3.2	519	2	0	0	0	9	4	2	2	3	0	9	4	3	0	14	0
3.3	629	9	6	1	0	11	1	3	5	15	0	21	4	3	1	11	2
3.4	963	10	3	0	1	10	3	10	12	8	0	21	2	6	3	15	5
3.5	502	7	3	2	2	3	0	6	3	3	0	8	1	2	0	13	1
4.1	546	4	3	2	3	4	4	2	7	7	1	8	1	2	1	8	2
4.2	379	3	0	1	0	3	1	3	1	4	2	8	0	2	0	9	1
4.3	1370	6	14	1	0	13	5	8	17	20	5	30	2	7	0	32	4
5.1	160	1	0	0	0	0	0	0	1	1	0	6	0	1	0	1	0
5.2	318	3	2	0	0	3	0	0	0	4	0	11	2	1	0	7	1
5.3	1968	23	6	2	1	19	8	10	21	20	5	51	7	7	3	44	11
Epi	62	1	0	0	0	3	0	1	0	0	0	1	0	1	0	2	0

Deme		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	3469	26	18	4	3	39	14	16	20	63	7	83	23	7	3	78	15
1.2	697	6	7	2	1	7	3	9	5	8	4	10	6	3	0	13	2
2.1	542	5	1	0	0	4	1	4	0	7	3	9	6	1	0	7	2
2.2	1029	10	9	0	1	11	3	7	5	12	2	31	1	3	0	29	3
2.3	1047	3	8	0	1	11	1	3	8	16	0	28	2	2	1	16	5
2.4	1708	12	12	1	1	11	1	12	12	21	2	40	8	8	1	45	4
2.5	298	0	4	0	0	2	0	3	1	4	1	12	2	1	0	5	2
3.1	250	3	0	0	0	1	1	0	1	1	1	3	0	0	3	1	0
3.2	1050	3	7	4	4	6	6	9	10	7	8	20	2	14	0	13	8
3.3	911	10	5	2	0	10	1	5	7	13	2	24	2	1	0	20	2
3.4	818	7	6	2	0	6	3	3	9	10	6	22	2	1	1	8	3
3.5	988	3	3	3	1	4	7	3	15	11	1	23	2	4	0	17	1
3.6	269	3	3	0	0	4	0	1	2	3	0	5	1	1	0	11	1
3.7	999	9	6	1	1	9	0	5	11	21	2	27	7	2	0	23	8
4.1	1466	10	11	2	2	10	7	9	6	21	3	39	12	10	0	35	6
4.2	1373	16	5	2	3	10	8	9	7	17	3	32	1	4	4	29	6
4.3	206	3	2	0	0	3	1	1	3	1	2	5	1	0	0	3	0
4.4	1758	9	10	6	2	22	6	15	24	19	2	53	1	5	1	39	9
4.5	866	8	2	3	1	8	3	8	12	12	2	13	0	3	1	15	0
4.6	317	0	2	0	0	5	0	1	3	3	1	11	0	2	0	8	0
4.7	106	0	0	0	0	2	0	0	0	1	0	3	0	1	0	3	0
4.8	1691	16	12	1	5	20	6	8	14	30	1	34	5	13	1	28	9
5.1	484	2	1	0	0	3	3	3	5	1	0	5	1	3	0	7	2
5.2	529	7	2	0	2	3	0	1	2	3	1	11	0	1	2	12	2
5.3	777	3	8	4	0	5	3	3	8	4	1	11	1	9	1	14	6
5.4	344	3	3	0	0	2	1	3	6	3	0	6	0	2	1	5	4
5.5	598	6	4	0	2	8	4	1	4	5	2	15	3	5	0	14	6

Prin		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	1157	10	7	2	0	21	6	6	1	22	4	33	2	10	5	15	9
1.2	852	6	12	2	0	2	1	7	3	14	1	15	2	1	3	25	6
1.3	2138	14	15	3	2	15	9	8	5	31	4	50	11	7	2	27	11
2.1	1296	6	6	1	0	9	2	14	3	16	2	19	2	4	1	38	7
2.2	619	13	4	1	0	3	3	5	0	8	1	28	2	5	1	11	3
2.3	772	14	9	2	0	11	4	5	5	4	3	25	3	4	0	30	2
2.4	447	2	3	0	1	8	0	0	1	14	0	13	1	1	0	6	0
2.5	263	2	1	0	0	0	0	2	0	5	0	13	1	1	0	11	0
2.6	1758	16	13	4	0	16	5	12	7	26	2	35	3	11	2	39	5
3.1	2445	9	21	4	0	18	12	19	7	42	6	52	15	12	5	39	9
3.2	1107	6	8	2	1	6	7	4	4	8	2	23	1	5	0	17	6
3.3	1485	3	11	2	2	11	2	10	4	15	2	22	8	5	0	24	4
4.1	710	4	5	0	0	6	0	1	7	12	0	31	10	4	3	14	2
4.2	1600	2	12	0	1	20	7	6	11	22	1	26	5	3	4	32	5
4.3	755	4	6	0	1	10	2	4	6	8	1	12	2	1	0	16	5
4.4	145	0	0	0	0	0	3	1	3	1	0	0	1	2	0	3	0
4.5	1271	6	14	1	1	11	4	4	10	21	2	18	1	6	2	31	7
5.1	722	7	13	3	1	7	5	7	6	7	1	14	4	9	0	16	0
5.2	1378	7	8	0	2	12	6	6	6	20	1	37	4	10	3	21	4
5.3	325	3	3	0	0	5	0	1	0	5	0	11	1	1	0	8	0
5.4	551	2	5	1	0	4	2	2	1	3	2	14	0	2	2	5	2
5.5	827	8	8	5	0	6	2	4	4	9	1	29	3	1	2	19	0

Pris		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	650	2	3	0	0	3	6	3	5	10	3	13	1	6	0	10	2
1.2	1940	14	12	3	0	16	6	10	11	25	1	36	5	2	5	32	11
1.3	2505	22	19	4	0	18	2	18	14	23	4	62	12	7	2	73	13
1.4	647	2	4	0	0	4	0	3	4	4	0	16	1	1	0	6	4
2.1	557	1	1	0	3	4	1	6	2	7	2	11	1	2	1	14	1
2.2	1050	5	5	0	2	5	2	8	4	18	2	19	0	4	0	21	3
2.3	391	1	1	1	0	2	0	3	3	3	1	4	0	1	0	3	0
2.4	794	8	6	1	0	9	3	4	2	23	1	34	2	1	2	15	1
2.5	41	0	2	0	0	0	0	0	0	1	0	1	0	0	0	0	0
2.6	1441	15	8	0	1	17	5	11	6	25	1	40	2	4	2	27	3
3.1	869	3	3	0	0	9	2	14	6	13	1	19	1	3	4	13	1
3.2	426	3	5	0	1	3	0	1	2	6	0	12	1	4	0	4	2
3.3	1527	8	4	0	5	14	3	8	11	26	1	29	3	9	2	17	6
3.4	522	2	4	0	2	5	2	1	1	4	1	7	0	5	0	9	1
3.5	1282	13	17	0	0	9	0	2	6	14	0	44	2	2	3	24	8
4.1	968	6	4	0	0	9	0	1	3	10	0	20	1	7	1	15	6
4.2	1371	18	4	3	0	11	4	3	7	28	1	34	1	5	7	18	3
4.3	449	6	5	0	0	3	3	1	6	11	1	10	4	2	0	11	0
4.4	506	3	3	1	4	0	0	7	2	9	1	10	0	1	2	8	0
4.5	2122	9	9	2	5	19	7	7	16	37	4	56	0	4	4	36	16
5.1	1450	9	6	0	1	11	15	5	11	12	0	26	0	10	0	32	9
5.2	589	7	0	1	0	6	2	1	1	7	0	17	0	1	1	9	6
5.3	368	6	2	1	0	2	1	1	4	3	2	4	2	2	0	13	0
5.4	818	6	3	1	0	8	2	3	5	8	1	15	1	1	1	18	6

Subj		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	693	5	2	0	2	3	1	1	2	13	1	29	0	8	2	17	0
1.2	1191	12	9	0	1	7	3	7	7	19	2	30	0	7	1	24	5
1.3	2379	16	9	4	1	22	8	24	17	28	1	63	8	14	5	60	10
1.4	309	6	2	0	0	2	1	2	2	3	0	11	1	3	0	1	0
1.5	848	8	5	0	0	4	3	2	16	9	0	25	0	5	3	16	1
2.1	3303	18	27	4	0	25	10	33	13	35	7	87	16	20	4	84	14
2.2	621	1	0	0	0	5	2	2	3	4	2	16	1	5	0	7	4
2.3	342	4	0	1	0	5	0	5	1	7	0	6	0	4	1	11	2
2.4	223	1	5	0	0	1	0	1	0	2	0	7	1	1	0	2	0
2.5	327	2	0	0	0	2	1	3	2	2	2	7	1	2	0	3	1
2.6	1135	19	4	2	1	6	2	9	8	11	2	24	0	7	0	27	7
3.1	220	3	1	0	2	3	1	5	6	3	0	3	1	2	0	6	0
3.2	1392	6	12	1	0	9	4	7	9	15	2	32	5	6	1	32	6
3.3	1104	7	5	1	1	9	4	8	6	8	3	10	0	10	0	18	3
3.4	737	1	15	1	0	4	2	6	1	6	0	13	4	5	0	25	2
3.5	457	4	4	0	0	7	0	3	2	5	0	17	1	1	0	9	1
3.6	1124	9	7	0	0	21	6	11	2	14	4	14	10	7	1	25	0
4.1	533	6	4	0	0	1	5	6	4	8	0	12	1	9	0	7	5
4.2	848	7	9	0	0	10	1	4	14	18	2	28	6	1	2	15	2
4.3	1479	9	17	4	1	10	0	8	8	13	2	29	5	14	2	37	5
4.4	518	0	2	1	0	6	0	3	3	4	1	13	3	2	1	3	1
4.5	1390	15	11	1	1	9	3	13	3	22	1	56	7	8	3	34	11
4.6	1034	11	12	0	2	6	2	2	2	6	2	24	2	3	1	27	2
5.1	272	1	3	0	1	2	1	1	1	2	0	8	2	3	0	8	3
5.2	1029	7	2	2	0	9	2	4	5	7	3	14	2	8	1	25	4
5.3	244	4	2	1	0	1	1	4	2	0	0	10	0	0	0	8	0
5.4	314	4	1	0	2	1	0	2	0	3	0	5	0	1	0	9	0
5.5	639	2	8	0	1	6	0	3	3	6	2	14	6	6	0	13	2
5.6	1074	11	4	2	0	5	4	11	6	10	0	20	3	4	0	32	6

Thom		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	1350	18	12	1	0	21	2	5	8	15	0	23	1	9	4	26	7
1.2	1259	4	1	1	1	12	6	12	6	12	3	15	2	8	1	14	7
1.3	1221	6	4	0	0	9	4	10	5	13	1	21	4	8	7	26	7
2.1	547	7	0	2	0	6	1	5	3	8	0	14	1	4	4	6	2
2.2	524	6	2	0	0	8	4	2	3	9	0	10	0	3	0	3	2
2.3	1323	11	4	0	2	9	3	4	6	8	1	42	2	5	2	31	7
2.4	153	1	0	0	0	1	3	5	1	0	0	7	0	0	0	3	2
2.5	957	7	5	0	0	5	4	9	7	17	1	14	0	7	1	17	6
3.1	3506	18	21	6	2	27	17	26	17	41	3	77	11	14	8	92	29
3.2	341	0	0	0	2	4	0	3	0	4	0	4	1	1	0	8	1
3.3	1246	13	7	0	2	13	3	13	7	15	1	28	1	9	1	22	2
4.1	634	9	2	0	5	3	2	2	1	9	1	20	1	3	0	14	3
4.2	1783	15	8	0	3	16	1	10	9	18	1	46	5	12	7	38	8
4.3	272	1	1	0	0	0	0	2	1	1	0	4	0	2	1	8	2
4.4	436	0	0	0	1	3	1	7	1	6	0	2	2	4	1	9	1
4.5	329	3	4	1	0	2	0	3	2	6	0	7	2	2	1	4	3

Thom (cont.)		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
5.1	620	1	2	0	0	11	1	2	2	3	0	6	1	2	1	8	4
5.2	402	4	4	0	0	6	2	2	3	2	2	5	0	2	0	6	1
5.3	227	2	0	0	0	1	0	2	4	2	0	3	0	1	1	6	1
5.4	66	1	0	0	0	0	0	0	1	1	1	2	0	1	0	2	0
5.5	570	1	2	0	0	5	1	2	8	5	1	14	0	2	0	11	0
5.6	652	3	6	1	1	4	4	6	11	8	0	16	1	2	1	11	3
5.7	152	3	0	0	0	3	0	2	0	1	0	2	1	0	1	3	2
5.8	45	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
5.9	239	3	3	0	0	3	1	3	4	3	0	10	0	0	0	3	3
5.10	449	2	2	0	2	3	3	4	3	4	0	7	2	2	1	8	2
5.11	460	1	3	0	5	4	3	2	3	4	1	13	0	2	0	8	1
5.12	1252	14	6	0	1	13	3	6	24	13	1	17	5	5	2	24	8
Vale		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
1.1	991	21	2	1	4	12	3	7	2	25	1	29	7	3	0	22	2
1.2	1521	13	9	1	0	8	5	10	3	24	1	51	1	9	9	34	10
1.3	2093	7	28	4	0	12	4	11	7	49	0	74	9	9	8	53	20
2.1	499	5	4	0	0	3	1	1	4	3	0	9	0	4	0	13	2
2.2	638	4	4	0	0	9	0	3	3	7	0	16	0	4	0	11	5
2.3	998	3	3	1	2	3	1	5	1	11	1	26	2	5	1	24	8
2.4	550	1	11	0	0	6	2	6	2	5	0	13	0	3	0	10	0
2.5	1003	9	15	1	1	9	1	9	4	8	0	34	2	5	0	19	6
2.6	344	2	7	1	1	2	0	0	1	6	0	16	1	1	1	4	3
3.1	3386	11	22	12	2	22	17	27	22	45	0	84	5	15	11	87	25
3.2	802	5	12	1	0	0	5	2	0	14	0	21	2	6	1	19	2
3.3	1563	14	10	4	3	11	15	7	9	23	1	30	3	2	0	37	12
4.1	1682	11	13	1	1	18	5	8	8	24	0	55	1	10	0	42	11
4.2	628	7	4	1	0	10	4	4	5	7	0	12	0	3	2	11	2
4.3	359	3	1	0	2	2	2	0	1	7	1	12	0	5	0	5	1
4.4	3165	23	19	7	2	22	13	17	23	46	1	69	0	7	3	67	22
5.1	298	0	2	0	0	6	1	2	0	5	0	7	0	0	0	10	1
5.2	1363	14	11	2	0	16	5	12	12	24	0	38	2	7	1	24	3
5.3	344	1	2	1	0	2	0	1	1	6	1	12	0	2	0	6	2
5.4	580	7	4	0	0	7	3	0	5	11	0	27	3	0	1	17	0
5.5	298	1	1	0	0	6	5	4	1	8	0	12	0	4	0	1	1
5.6	490	3	2	0	1	8	2	1	2	14	0	21	1	5	0	10	5
5.7	112	1	0	0	0	2	2	0	0	0	1	6	0	2	0	6	1
5.8	965	5	4	1	2	7	3	2	4	24	0	33	2	1	0	22	4
Epi	172	1	1	0	0	2	0	0	0	0	0	7	0	0	1	3	2

H8		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
Pro	268	0	1	0	0	3	0	1	3	6	0	8	0	1	0	3	0
1.1	1868	10	2	0	10	30	2	7	6	42	1	94	4	1	9	25	22
1.2	1742	5	8	0	3	31	5	7	2	43	0	78	5	4	9	27	23
1.3	587	3	7	0	0	3	1	4	4	16	2	18	1	2	1	11	2
1.4	941	9	11	0	0	8	3	2	7	13	0	15	6	2	2	23	5
2.1	1439	14	5	2	2	14	6	3	8	29	2	53	1	5	4	33	9
2.2	1220	12	8	2	0	12	3	2	4	29	1	41	5	5	3	10	9
2.3	898	8	7	0	0	10	1	8	3	18	0	21	0	4	5	17	5
2.4	1924	6	15	0	10	31	3	8	5	52	0	64	5	3	9	27	13
3.1	1525	10	9	3	1	14	2	8	6	19	0	32	2	3	1	25	8
3.2a	1663	8	5	0	9	31	2	11	7	39	1	81	1	1	9	25	11
3.2b	2185	24	6	7	1	28	7	8	12	55	1	85	4	4	3	26	22
4.1	999	11	9	0	1	11	1	5	2	35	1	69	3	1	5	18	11
4.2	1431	6	10	1	1	20	2	7	9	20	1	38	3	1	6	25	15
5.1	1507	1	8	2	2	17	6	6	9	34	0	51	2	3	7	16	8
5.2	296	1	0	0	0	0	2	0	0	6	1	9	0	2	0	7	1
5.3	1550	12	18	3	1	21	5	6	6	39	2	38	1	6	6	25	9
5.4	807	5	12	0	2	10	1	4	1	15	0	40	7	1	2	23	7
5.5	653	11	0	0	1	7	5	3	4	12	0	20	0	1	1	17	5
Epi	132	3	3	0	0	1	0	0	0	1	0	4	0	0	1	3	1

TNK		al	ar	dr	dd	in	mu	no	nw	of	su	th	ts	to	wh	F	S
Pro	273	1	2	0	0	1	1	0	0	5	1	7	0	1	0	5	3
1.1	1821	8	5	1	1	21	3	6	8	43	0	64	1	2	10	31	23
1.2	954	1	3	0	3	22	3	2	3	30	0	42	1	2	5	12	7
1.3	804	2	3	1	3	19	2	5	2	15	2	19	2	1	4	12	7
1.4	413	3	4	0	0	10	0	1	0	12	0	11	0	0	2	13	8
1.5	108	1	1	0	0	0	0	0	0	1	1	3	0	0	0	1	1
2.1	497	1	6	0	1	7	0	5	1	20	0	20	1	0	0	9	3
2.2	2402	18	16	3	0	19	17	12	12	44	4	58	4	5	2	40	17
2.3	744	6	3	1	0	10	4	3	2	8	1	30	0	2	0	10	2
2.4	288	0	2	0	0	5	0	0	0	2	0	3	1	1	0	7	0
2.5	573	5	13	2	0	4	2	0	0	6	2	5	1	2	0	9	1
2.6	355	4	1	0	1	0	0	2	1	3	1	8	0	1	0	11	1
3.1	1051	3	2	1	0	17	2	1	3	28	1	31	5	1	4	12	4
3.2	343	3	2	0	0	5	0	4	2	4	0	12	1	0	1	3	1
3.3	502	3	4	0	4	5	1	6	5	3	0	9	1	3	0	6	2
3.4	250	4	2	0	0	2	0	1	8	3	0	8	0	1	0	3	2
3.5	1241	9	11	0	1	7	1	5	9	14	1	50	0	1	1	11	8
3.6	2717	22	19	5	2	29	7	19	10	34	1	46	11	8	4	54	9
4.1	1353	10	10	0	4	9	9	6	4	23	0	42	2	2	2	22	6
4.2	1349	9	10	1	0	14	8	7	9	31	0	38	5	2	4	34	10
4.3	877	3	4	0	2	19	2	2	4	22	0	25	0	2	4	9	8
5.1	1392	3	1	0	2	16	6	3	2	40	0	54	2	2	9	15	9
5.2	1039	5	7	0	3	12	4	3	2	13	0	26	2	5	0	15	1
5.3	1211	4	7	0	6	13	8	7	5	24	1	56	1	2	6	25	9
5.4	1158	5	8	0	5	9	0	0	3	22	1	44	0	3	6	19	14
Epi	169	1	0	1	0	1	0	3	2	0	0	3	0	0	0	2	2

Bibliography

[1] Peter Alexander. Conjectural History, or *Henry VIII. Essays and Studies,* XVI:85–120, 1939.

[2] William W. Appleton. *Beaumont and Fletcher: A Critical Study.* George Allen & Unwin, London, 1956.

[3] Stephen F. Austin. *A Computer-aided Technique for Stylistic Discrimination: The Authorship of "Greene's Groatworth of Wit".* Austin State College, Nagodoches, Texas, April 1969. Sponsored by U.S. Dept. of Health, Education and Welfare, Office of Education, Bureau of Research.

[4] Norman T. J. Bailey. *Statistical Methods in Biology.* Hodder and Stoughton, London, 2nd edition, 1981.

[5] W. M. Baillie. Authorship Attribution in Jacobean Dramatic Texts. In J. L. Mitchell, editor, *Computers in the Humanities,* pages 73–81, Edinburgh University Press, 1974.

[6] R. C. Bald. *Bibliographical Studies in the Beaumont and Fletcher Folio of 1647.* The Bibliographical Society, 1938.

[7] Charles Barber. *Early Modern English. The Language Library,* André Deutsch, London, 1976.

David Bevington, editor. *The Complete Works of William Shakespeare.* See Shakespeare [126].

[8] Francis Beaumont and John Fletcher. *The Dramatic Works in the Beaumont and Fletcher Canon.* Fredson Bowers, editor. Volume 1, Cambridge University Press, 1961.

[9] Francis Beaumont and John Fletcher. *The Dramatic Works in the Beaumont and Fletcher Canon.* Fredson Bowers, editor. Volume 3, Cambridge University Press, 1973.

[10] Francis Beaumont and John Fletcher. *The Dramatic Works in the Beaumont and Fletcher Canon.* Fredson Bowers, editor. Volume 4, Cambridge University Press, 1979.

[11] Francis Beaumont and John Fletcher. *The Dramatic Works in the Beaumont and Fletcher Canon.* Fredson Bowers, editor. Volume 5, Cambridge University Press, 1982.

[12] G. E. Bentley. *The Jacobean and Caroline Stage.* 5 volumes, The Clarendon Press, Oxford, 1941–1956.

[13] C. I. Bliss. Fitting the Negative Binomial Distribution to Biological Data. *Biometrics,* 9:176–200, June 1953.

[14] Norbert Bolz. Are Robert Greene's "Autobiographies" Fakes? *The Shakespeare Newsletter,* XXIX:6(161):43, Dec 1979.

Fredson Bowers, editor. *The Dramatic Works in the Beaumont and Fletcher Canon.* See Beaumont and Fletcher [8,9,35,36].

[15] Fredson Bowers. *On Editing Shakespeare and the Elizabethan Dramatists.* University of Pennsylvania Library, 1935.

[16] Barron Brainerd. The Chronology of Shakespeare's Plays: A Statistical Study. *Computers and the Humanities,* 14:221–230, 1980.

[17] Barron Brainerd. On the Distinction between a Novel and a Romance: A Discriminant Analysis. *Computers and the Humanities,* 7:259–270, 1973.

[18] Barron Brainerd. Pronouns and Genre in Shakespeare's Drama. *Computers and the Humanities,* 13:3–16, 1979.

[19] Barron Brainerd. Statistical Analysis of Lexical Data Using Chi-squared and Related Distributions. *Computers and the Humanities,* 9:161–178, 1975.

[20] Claude S. Brinegar. Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship. *Journal of the American Statistical Association,* March 1963.

[21] T. A. Brown and J. Koplowitz. The Weighted Nearest Neighbor Rule for Class Dependent Sample Sizes. *IEEE Transactions on Information Theory,* IT-25:671–619, 1979.

[22] J. F. Burrows. Modal Verbs and Moral Principles: An Aspect of Jane Austen's Style. *Literary and Linguistic Computing,* 1(1):9–23, 1986.

[23] Edmund K. Chambers. *The Elizabethan Stage.* 4 Volumes, The Clarendon Press, Oxford, 1923.

[24] T. M. Cover and P. E. Hart. Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory,* IT-13(1), January 1967.

[25] D. R. Cox and L. Brandwood. On a Discriminatory Problem Connected with the Works of Plato. *Journal of the Royal Statistical Society B,* 21:195–200, 1959.

[26] Fred J. Damerau. The Use of Function Word Frequencies as Indicators of Style. *Computers and the Humanities,* 9:271–280, 1975.

[27] Bruno de Finetti. *Theory of Probability.* Volume 1, John Wiley & Sons, London, 1974.

[28] Pierre A. Devijver and Josef Kittler. *Pattern Recognition: A Statistical Approach.* Prentice-Hall International, London, 1982.

[29] L. P. Devroye and T. J. Wagner. Distribution Free Inequalities for the Deleted and Holdout Error Estimates. *IEEE Transactions on Information Theory,* IT-25:202–207, 1979.

[30] Philip Edwards. On the Design of "The Two Noble Kinsmen". In *The Two Noble Kinsmen* [39], Clifford Leech, editor, pages 243–261. From *A Review of Modern Literature,* 5:89–105, 1964.

[31] Alvar Ellegård. *A Statistical Method for Determining Authorship: The Junius Letters 1769–1772. Gothenberg Studies in English,* University of Gothenberg Press, 1962.

[32] Alvar Ellegård. *Who was Junius?* Almquist and Wiksell, Stockholm, 1962.

[33] David V. Erdman and Ephim G. Fogel, editors. *Evidence for Authorship: Essays on Problems of Attribution.* Cornell University Press, 1966.

G. B. Evans, editor. *The Riverside Shakespeare.* See Shakespeare [136].

[34] G. Blakemore Evans. Shakespeare's Text: Approaches and Problems. In Kenneth Muir and S. Schoenbaum, editors, *A New Companion to Shakespeare Studies,* pages 222–238, Cambridge University Press, 1971.

[35] Willard E. Farnham. Colloquial Contractions in Beaumont, Fletcher, Massinger and Shakespeare as a Test of Authorship. *Publications of the Modern Language Association,* 31:326–358, 1916.

[36] Sir Ronald A. Fisher. *Statistical Methods for Research Workers.* Oliver & Boyd, Edinburgh, 13th edition, 1958.

[37] John Fletcher and William Shakespeare. *The Two Noble Kinsmen.* Harold Littledale, editor. Publications of the New Shakspere Society, 1876 (Part 1) and 1885 (Part 2).

[38] John Fletcher and William Shakespeare. *The Two Noble Kinsmen.* G. R. Proudfoot, editor. *Regents Renaissance Drama Series,* Edward Arnold, 1970.

[39] John Fletcher and William Shakespeare. *The Two Noble Kinsmen.* Clifford Leech, editor. *The Signet Class Shakespeare,* The New American Library, New York, 1966.

R. A. Foakes, editor. *King Henry VIII.* See Shakespeare [130].

[40] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, New York, 1972.

[41] F. J. Furnivall. Mr. Hickson's Divison of *The Two Noble Kinsmen,* Confirmed by the Stopt-Line Test. In *Transactions of the New Shakspere Society,* Appendix, pages 64*–65*, London, 1874.

[42] M. Goldstein. Comparison of Some Density Estimate Classification Procedures. *Journal of the American Statistical Society,* 70:666–669, 1975.

[43] W. W. Greg. *The Shakespeare First Folio: Its Bibliographic and Textual History.* The Clarendon Press, Oxford, 1955.

[44] J. D. F. Habbema, J. Hermans, and K. van den Broek. A Stepwise Discriminant Analysis Program Using Density Estimation. In G. Bruckman, editor, *Compstat 1974,* pages 101–110, Physica-Verlag, Vienna, 1974.

Antony Hammond, editor. *King Richard III.* See Shakespeare [133].

[45] D. J. Hand. *Discrimination and Classification.* John Wiley & Sons, Chichester, 1981.

[46] Alfred Hart. *Shakespeare and the Vocabulary of The Two Noble Kinsmen,* pages 242–256. Melbourne University Press, 1934.

[47] J. Hermans and J. D. F. Habbema. The ALLOC Package: Multigroup Discriminant Analysis Programs Based on Direct Density Estimation. In Johannes Gordesch and Peter Naeve, editors, *Compstat 1976,* pages 350–357, Physica-Verlag, Vienna, 1976.

[48] Samuel Hickson. The Shares of Shakspere and Fletcher in *The Two Noble Kinsmen.* In *Transactions of the New Shakspere Society,* Appendix, pages 25*–61*, London, 1874. First printed in *Westminster Review,* XLVII (April 1847), pages 57ff.

Charlton Hinman, editor. *The Norton Facsimile: The First Folio of Shakespeare.* See Shakespeare [134].

[49] Charlton Hinman. *The Printing and Proof-Reading of the First Folio of Shakespeare.* 2 Volumes, The Clarendon Press, Oxford, 1963.

[50] Susan Hockey. *A Guide to Computer Applications in the Humanities.* Duckworth, London, 1980.

[51] D. I. Holmes. The Analysis of Literary Style — A Review. *Journal of the Royal Statistical Society A,* 148:328–341, 1985.

[52] T. H. Howard-Hill. The Oxford Old-Spelling Shakespeare Concordances. *Studies in Bibliography,* 22:143–164, 1969.

[53] Cyrus Hoy. Fletcherian Romantic Comedy. *Research Opportunities in Renaissance Drama,* XXVII:3–11, 1984.

[54] Cyrus Hoy. The Shares of Fletcher and his Collaborators in the Beaumont and Fletcher Canon (I). *Studies in Bibliography,* VIII:129–146, 1956.

[55] Cyrus Hoy. The Shares of Fletcher and his Collaborators in the Beaumont and Fletcher Canon (VII). *Studies in Bibliography,* XV:71–90, 1962.

A. R. Humphries, editor. *King Henry the Eighth.* See Shakespeare [129].

A. R. Humphries, editor. *The Second Part of King Henry the Fourth.* See Shakespeare [137].

[56] MacD. P. Jackson. A Hint for Investigators of Authorship. *The Shakespeare Newsletter,* XXIX:6(161):43–44, Dec 1979.

[57] Harold Jeffreys. *Scientific Inference.* Cambridge University Press, 1931.

Harold Jenkins, editor. *Hamlet.* See Shakespeare [128].

[58] K. W. Kemp. Aspects of the Statistical Analysis and Effective Use of Linguistic Data. *Association for Literary and Linguistic Computing Bulletin,* 1(1):14–22, 1976.

[59] Kenneth Kemp. Personal Observations on the Use of Statistical Methods in Quantitative Linguistics. In Alan Jones and R. F. Churchhouse, editors, *The Computer in Literary and Linguistic Studies,* pages 59–77, The University of Wales Press, Cardiff, 1976.

[60] Maurice G. Kendall. *The Advanced Theory of Statistics.* Volume 2, Charles Griffin & Company, London, 2nd edition, 1948.

[61] Anthony Kenney. *The Computation of Style: An Introduction to Statistics for Students of Literature and Humanities.* Pergamon Press, Oxford, 1982.

[62] John Maynard Keynes. *A Treatise on Probability.* Macmillan and Company, London, 1929.

[63] Geir Kjetsaa. Written by Dostoevsky? *Association for Literary and Linguistic Computing Journal,* 2:25–33, 1981.

[64] G. Wilson Knight. A Note on *Henry VIII.* In *The Famous History of the Life of King Henry the Eighth* [127], S. Schoenbaum, editor, pages 217–225. From *Shakespeare and Religion.* Barnes & Noble, New York, 1967.

[65] Peter A. Lachenbruch. *Discriminant Analysis.* Hafner Press, New York, 1975.

[66] Wayne A. Larsen and Alvin C. Rencher. Who Wrote the Book of Mormon? An Analysis of Wordprints. In Noel B. Reynolds, editor, *Book of Mormon Authorship,* Brigham Young Religious Studies Center, 1983.

[67] Robert A. Law. The Double Authorship of *Henry VIII. Studies in Philology,* LVI:471–486, 1959.

[68] G. R. Ledger. A New Approach to Stylometry. *Association for Literary and Linguistic Computing Bulletin,* 13(3):67–72, 1985.

[69] Clifford Leech. *The John Fletcher Plays.* Chatto and Windus, London, 1962.

Clifford Leech, editor. *The Two Noble Kinsmen.* See Shakespeare [39].

[70] M. Levison, A. Q. Morton, and W. C. Wake. On Certain Statistical Features of the Pauline Epistles. *The Philosophical Journal*, 3(2):129–148, 1966.

Harold Littledale, editor. *The Two Noble Kinsmen.* See Fletcher and Shakespeare [37].

[71] Louis Marder. A Guide to 50 Computer Projects in Shakespeare. *The Shakespeare Newsletter*, 15:52, December 1965.

[72] Louis Marder. The New Disintegration or Reintegration of the Shakespeare Canon. *The Shakespeare Newsletter*, XXXII:2–3(174):2, Apr–May 1982.

[73] Louis Marder. Scholars Dispute Shakespeare Data. *The Shakespeare Newsletter*, 37–38, Winter 1983.

[74] Louis Marder. Stylometric Analysis and the Pericles Problem. *The Shakespeare Newsletter*, XXVI:6:46, Dec 1976.

[75] Louis Marder. Stylometrics: The New Authorship Weapon. *The Shakespeare Newsletter*, XXIX:6(161):41–42, Dec 1979.

[76] Louis Marder. Stylometry: Possibilities and Problems. *The Shakespeare Newsletter*, XXXIV:1(181):4, Spring 1984.

[77] Louis Marder. Stylometry "Proves" Entire "Sir Thomas More" is All Shakespeare's. *The Shakespeare Newsletter*, XXX:4(165):29–30, Sep 1980.

[78] Louis Marder. Stylometry: The Controversy Continues. *The Shakespeare Newsletter*, XXXIV:3(183):28, Fall 1984.

[79] Ian Marshall. Choice of Grammatical Word-Class without Global Syntactic Analysis: Tagging Words in the LOB Corpus. *Computers and the Humanities*, 17(3):139–150, 1983.

J. C. Maxwell, editor. *King Henry VIII.* See Shakespeare [131].

[80] William McColly and Dennis Weier. Literary Attribution and Likelihood-Ratio Tests: The Case of the Middle English *Pearl*-Poems. *Computers and the Humanities*, 17(2):65–75, June 1983.

[81] R. J. McKay and N. A. Campbell. Variable Selection Techniques in Discriminant Analysis: I. Description II. Allocation. *British Journal of Mathematical and Statistical Psychology*, 35:1–41, 1982.

[82] Ronald B. McKerrow. *An Introduction to Bibliography for Literary Students.* The Clarendon Press, Oxford, 1928.

[83] Ronald B. McKerrow. *Prolegomena for the Oxford Shakespeare: A Study in Editorial Method.* The Clarendon Press, Oxford, 1939.

[84] T. C. Mendenhall. The Characteristic Curves of Composition. *Science,* 11(214):237–249, 11 March 1887.

[85] T. C. Mendenhall. A Mechanical Solution of a Literary Problem. *Popular Science Monthly,* 60(7):97–105, 1901.

[86] Thomas Merriam. The Authorship of Sir Thomas More. *Association for Literary and Linguistic Computing Bulletin,* 10(1):1–7, 1982.

[87] Thomas Merriam. Did Shakespeare Write *Sir Thomas More*? *The Shakespeare Newsletter,* XXXI:1(168):6, Feb 1981.

[88] Thomas Merriam. 'Henry VIII' and the Integrity of the First Folio. *The Bard,* 3:69–73, 1981.

[89] Thomas Merriam. What Shakespeare Wrote in *Henry VIII*: Part One. *The Bard,* 2(3):95–99, 1979.

[90] Thomas Merriam. What Shakespeare Wrote in *Henry VIII*: Part Two. *The Bard,* 2(4):111–118, 1980.

[91] Thomas Merriam and M. W. A. Smith. The Authorship Controversy of Sir Thomas More. *Literary and Linguistic Computing,* 1(2):104–108, 1986. Merriam's response to Smith's earlier criticism and Smith's further response.

[92] Thomas van Ness Merriam. *The Consonance of Literary Elements with Mathematical Models: A Study of Authorship in the Huntingdon Plays.* Master's thesis, University of London King's College, 1985.

[93] Tom Merriam. Morton vs Smith: An Objective Analysis. *The Shakespeare Newsletter,* XXXIV:1(181):5, Spring 1984.

[94] G. Harold Metz. Disputed Shakespearean Texts and Stylometric Analysis. In D. C. Greetham and W. Speed Hill, editors, *Text,* pages 149–171, The Society for Textual Scholarship, AMS Press, New York, 1985. Number 2 of a series published annually.

[95] G. Harold Metz. Stylometric Analysis and *Sir Thomas More. The Shakespeare Newsletter,* XXXI:1(168):6, Feb 1981.

[96] G. Harold Metz. A Stylometric Comparison of Shakespeare's *Titus Andronicus, Pericles* and *Julius Caesar*. *The Shakespeare Newsletter,* XXIX:6(161):42, Dec 1979.

[97] S. Michaelson and A. Q. Morton. Last Words. *New Testament Studies,* 18:192–208, 1972.

[98] S. Michaelson and A. Q. Morton. The New Stylometry: A One-word Test of Authorship for Greek Writers. *The Classical Quarterly,* 22(1):89–102, May 1972.

[99] S. Michaelson and A. Q. Morton. Positional Stylometry. In A. J. Aitken, R. W. Bailey, and N. Hamilton-Smith, editors, *The Computer and Literary Studies,* pages 69–84, Edinburgh University Press, 1973.

[100] S. Michaelson and A. Q. Morton. The Spaces in Between: A Multiple Test of Authorship for Greek Writers. In *L.A.S.L.A. Revue,* International Organization for Ancient Languages Analysis by Computer, Liège (Belgium), 1972. Number 1.

[101] S. Michaelson, A. Q. Morton, and N. Hamilton-Smith. *Justice for Helander.* Technical Report CSR–42–79, University of Edinburgh Department of Computer Science, June 1979.

[102] S. Michaelson, A. Q. Morton, and N. Hamilton-Smith. *To Couple Is the Custom.* Technical Report CSR–22–79, University of Edinburgh Department of Computer Science, June 1979.

[103] Sidney Michaelson and Andrew Morton. Things Aint What They Used to Be. In Alan Jones and R. F. Churchhouse, editors, *The Computer in Literary and Linguistic Studies,* pages 78–84, The University of Wales Press, Cardiff, 1976.

[104] Marco Mincoff. The Authorship of *The Two Noble Kinsmen. English Studies,* XXXIII:97–115, 1952.

[105] Marco Mincoff. *Henry VIII* and Fletcher. *Shakespeare Quarterly,* XII:239–260, 1961.

[106] A. Q. Morton. The Authorship of Greek Prose. *Journal of the Royal Statistical Society A,* 128:169–224, 1965.

[107] A. Q. Morton. Fingerprinting the Mind. Text of a lecture delivered to the Royal Society of Edinburgh on 4 February 1985; provided by the author.

[108] A. Q. Morton. *Literary Detection.* Bowker, 1978.

[109] A. Q. Morton. Once: A Test of Authorship Based on Words which Are Not Repeated in the Sample. *Literary and Linguistic Computing,* 1(1):1–8, 1986.

[110] A. Q. Morton and J. McLeman. *The Genesis of John.* The Saint Andrew Press, Edinburgh, 1980.

[111] A. Q. Morton and James McLeman. *Paul, the Man and the Myth: A Study in the Authorship of Greek Prose.* Hodder and Stoughton, London, 1966.

[112] A. Q. Morton and S. Michaelson. *The Nature of Stylometry.* Technical Report, University of Edinburgh Department of Computer Science, 1982. Prepared for *Stylometrics '82* workshop.

[113] Frederick Mosteller and David L. Wallace. *Inference and Disputed Authorship: The Federalist.* Addison-Wesley, Reading, Massachusetts, 1964.

[114] Kenneth Muir. *Shakespeare as Collaborator.* Methuen, London, 1960.

[115] Kenneth Muir. Shakespeare's Hand in *The Two Noble Kinsmen. Shakespeare Survey,* 11:50–59, 1958.

[116] Robert L. Oakman. *Computer Methods for Literary Research.* University of South Carolina Press, 1980.

[117] D. P. O'Brien and A. C. Darnell. *Authorship Puzzles in the History of Economics: A Statistical Approach.* The Macmillan Press, London, 1982.

[118] D. P. O'Brien, A. C. Darnell, and J. Peters. A Statistical Technique for the Investigation of Authorship Puzzles. May 1983. Working Paper No. 59; privately supplied by O'Brien.

[119] Ants Oras. "Extra Monosyllables" in *Henry VIII. Journal of English and Germanic Philology,* 52:198–213, 1953.

[120] A. C. Partridge. *Orthography in Shakespeare and Elizabethan Drama: A Study of Colloquial Contractions, Elision, Prosody and Punctuation.* Edward Arnold, London, 1964.

[121] F. C. Powell. *Statistical Tables for the Social, Biological and Physical Sciences.* Cambridge University Press, 1982.

G. R. Proudfoot, editor. *The Two Noble Kinsmen.* See Fletcher and Shakespeare [38].

[122] J. Remme, J. D. F. Habbema, and J. Hermans. A Simulative Comparison of Linear, Quadratic and Kernel Discrimination. *Journal of Statistical Computations and Simulation*, 11:87–106, 1980.

Norman Sanders, editor. *Othello*. See Shakespeare [135].

[123] *SAS User's Guide: Statistics*, 1982 edition. SAS Institute Inc., Cary, North Carolina, 1982.

[124] Leonard J. Savage. *The Foundation of Statistics*. John Wiley & Sons, New York, 1954.

S. Schoenbaum, editor. *The Famous History of the Life of King Henry the Eighth*. See Shakespeare [127].

[125] S. Schoenbaum. *Internal Evidence and Elizabethan Dramatic Authorship: An Essay in Literary History and Method*. Edward Arnold, London, 1966.

[126] William Shakespeare. *The Complete Works*. David Bevington, editor. Scott, Foresman and Company, Glenview, Illinois, 1980.

[127] William Shakespeare. *The Famous History of the Life of King Henry the Eighth*. S. Schoenbaum, editor. *The Signet Class Shakespeare*, The New American Library, New York, 1967.

[128] William Shakespeare. *Hamlet*. Harold Jenkins, editor. *The Arden Shakespeare*, Methuen, London, 1982.

[129] William Shakespeare. *King Henry the Eighth*. A. R. Humphries, editor. *New Penguin Shakespeare*, Penguin, Hammondsworth, England, 1971.

[130] William Shakespeare. *King Henry VIII*. R. A. Foakes, editor. *The Arden Shakespeare*, Methuen, London, 3rd edition, 1957.

[131] William Shakespeare. *King Henry VIII*. J. C. Maxwell, editor. Cambridge University Press, 1962.

[132] William Shakespeare. *King Richard II*. Peter Ure, editor. *The Arden Shakespeare*, Methuen, London, 5th edition, 1961.

[133] William Shakespeare. *King Richard III*. Antony Hammond, editor. *The Arden Shakespeare*, Methuen, London, 1981.

[134] William Shakespeare. *The Norton Facsimile: The First Folio of Shakespeare*. Charlton Hinman, editor. Paul Hamlyn, London, 1968. Photographic facsimile of 1623 Folio and introduction.

[135] William Shakespeare. *Othello.* Norman Sanders, editor. *The New Cambridge Shakespeare,* The Cambridge University Press, 1984.

[136] William Shakespeare. *The Riverside Shakespeare.* G. B. Evans, editor. Houghton Mifflin, Boston, 1974.

[137] William Shakespeare. *The Second Part of King Henry the Fourth.* A. R. Humphries, editor. *The Arden Shakespeare,* Methuen, London, 1966.

[138] William Shakespeare. *Shakespeare's Plays in Quarto: a Facsimile Edition of Copies Primarily from the Henry E. Huntingdon Library.* M. J. B. Allen and K. Muir, editors. University of California Press, Berkeley, 1981.

[139] G. B. Shand and Raymond C. Shady, editors. *Play-Texts in Old Spelling: Papers from the Gleneldon Conference. AMS Studies in the Renaissance,* AMS Press, New York, 1984.

[140] Alexander [sic] Q. Morton. Stylometry vs "Stylometry". *The Shakespeare Newsletter,* XXXIV:1(181):5, Spring 1984.

[141] H. S. Sichel. On a Distribution Representing Sentence-Length in Written Prose. *Journal of the Royal Statistical Society A,* 137:25–34, 1974.

[142] B. W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London, 1986.

[143] M. W. A. Smith. The Authorship of "A Lover's Complaint": An Application of Statistical Stylometry to Poetry. *Computers and the Humanities,* 18(1):23–38, Jan.–Mar. 1984.

[144] M. W. A. Smith. The Authorship of *Pericles:* An Initial Investigation. *The Bard,* 3(4):143–176, 1982.

[145] M. W. A. Smith. Critical Reflections on the Determination of Authorship by Statistics: Part 1. Shakespeare, Bacon and Marlowe. *The Shakespeare Newsletter,* XXXIV:1(181):4–5, Spring 1984.

[146] M. W. A. Smith. Critical Reflections on the Determination of Authorship by Statistics: Part 2. Morton, Merriam and *Pericles. The Shakespeare Newsletter,* XXXIV:3(183):28–33, Fall 1984.

[147] M. W. A. Smith. An Initial Investigation of the Authorship of *Pericles. The Shakespeare Newsletter,* 32, Fall 1983.

[148] M. W. A. Smith. An Investigation of Morton's Method to Distinguish Elizabethan Playwrights. *Computers and the Humanities,* 19:3–21, 1985.

[149] M. W. A. Smith. Recent Experience and New Developments of Methods for the Determination of Authorship. *Association for Literary and Linguistic Computing Bulletin*, 11(3):73–82, 1983.

[150] M. W. A. Smith. A Stylometric Analysis of *Hero and Leander*. *The Bard*, 3:105–132, 1982.

[151] George W. Snedecor and William G. Cochran. *Statistical Methods*. Iowa State University Press, Ames, 7th edition, 1980.

[152] James Spedding. On the Several Shares of Shakspere and Fletcher in the Play of *Henry VIII*. In *Transactions of the New Shakspere Society*, Appendix, pages 1*–18*, London, 1874. First printed as "Who Wrote Shakspeare's *Henry VIII*?," *The Gentleman's Magazine*, n.s., XXXIV (1850), pp. 115–123.

[153] Theodore Spencer. *The Two Noble Kinsmen*. In *The Two Noble Kinsmen* [39], Clifford Leech, editor, pages 217–242. From *Modern Philology*, XXXVI:255–276, 1939.

[154] M. Spevack, H. J. Neuhaus, and T. Finkenstaedt. SHAD: A Shakespeare Dictionary. In J. L. Mitchell, editor, *Computers in the Humanities*, pages 111–123, Edinburgh University Press, 1974.

[155] Marvin Spevack. *A Complete and Systematic Concordance to the Works of Shakespeare*. Georg Olms, Hildesheim, 1968.

[156] Arthur Colby Sprague. *Shakespeare's Histories: Plays for the Stage*. The Society for Theatre Research, London, 1964.

[157] Caroline F. E. Spurgeon. *Shakespeare's Imagery and What It Tells Us*. Cambridge University Press, 1935.

[158] David Tallentire. Confirming Intuitions about Style, Using Concordances. In Alan Jones and R. F. Churchhouse, editors, *The Computer in Literary and Linguistic Studies*, pages 309–328, The University of Wales Press, Cardiff, 1976.

[159] David Tallentire. Towards an Archive of Lexical Norms: A Proposal. In A. J. Aitken, R. W. Bailey, and N. Hamilton-Smith, editors, *The Computer and Literary Studies*, pages 39–60, Edinburgh University Press, 1973.

[160] Gary Taylor. Three Studies in the Text of *Henry V*. In *Modernizing Shakespeare's Spelling* [169], The Oxford University Press, 1979.

[161] N. D. Thomson. Literary Statistics. Six parts. Association for Literary and Linguistic Computing Bulletin Vol. 1, No. 3: 10–14; Vol. 2, No. 1: 10–15; Vol. 2, No. 2: 42–47; Vol. 2, No. 3: 55–61; Vol. 3, No. 1: 29–35; Vol. 3, No. 2: 166–171.

[162] Louis Ule. *A Concordance to the Works of Christopher Marlowe.* Georg Olms Verlag, Hildesheim, 1979.

[163] Louis Ule. Recent Progress in Computer Methods of Authorship Determination. *Association for Literary and Linguistic Computing Bulletin,* 10(3):73–89, 1982.

Peter Ure, editor. *King Richard II.* See Shakespeare [132].

[164] Mary-Clair van Leunen. *A Handbook for Scholars.* Alfred A. Knopf, New York, 1979.

[165] Richard von Mises. *Mathematical Theory of Probability and Statistics.* Hilda Geiringer, editor. Academic Press, New York, 1964.

[166] W. C. Wake. Sentence-Length Distributions of Greek Authors. *Journal of the Royal Statistical Society A,* 120:331–346, 1957.

[167] T. R. Waldo. Review of Austin's "A Computer-aided Technique for Stylistic Discrimination..." [3]. *Computers and the Humanities,* 7(2):109–110, November 1972.

[168] J. K. Walton. *The Quarto Copy for the First Folio of Shakespeare.* Dublin University Press, 1971.

[169] Stanley Wells. *Modernizing Shakespeare's Spelling.* The Clarendon Press, Oxford, 1979.

[170] Stanley Wells. *Re-editing Shakespeare for the Modern Reader.* The Clarendon Press, Oxford, 1984.

[171] D. Wickmann. On Disputed Authorship, Statistically. *Association for Literary and Linguistic Computing Bulletin,* 4(1):32–41, 1976.

[172] C. B. Williams. *Style and Vocabulary: Numerical Studies.* Griffin, London, 1970.