

THE ATTRIBUTION OF AUTHORSHIP:  
A COMPUTERIZED METHOD EVALUATED AND COMPARED  
WITH OTHER METHODS PAST AND FUTURE

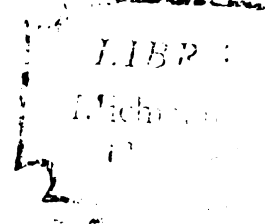
THESIS FOR THE DEGREE OF PH. D.

MICHIGAN STATE UNIVERSITY

GEORGE W. ZIMMER

1968

THESIS



This is to certify that the

thesis entitled

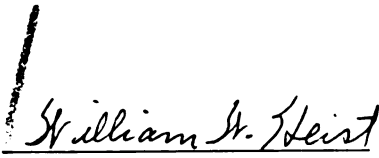
THE ATTRIBUTION OF AUTHORSHIP:  
A COMPUTERIZED METHOD EVALUATED AND COMPARED  
WITH OTHER METHODS PAST AND FUTURE

presented by

George W. Zimmer

has been accepted towards fulfillment  
of the requirements for

Ph. D. degree in English

  
Major professor

Date April 16, 1968

C-030

5/10/97.

## ABSTRACT

### THE ATTRIBUTION OF AUTHORSHIP: A COMPUTERIZED METHOD EVALUATED AND COMPARED WITH OTHER METHODS PAST AND FUTURE

By

George W. Zimmer

The proving of authorship by statistical means has a long but inglorious history in the field of English scholarship. What usually has happened is that an undisinterested scholar, out to "prove" that (for example) there was a "Pearl-poet" to whom can be attributed three or four additional Middle English poems, lists elements that the anonymous Pearl has in common with the other poems, and concludes that on the basis of his "statistics" the poems must have been by the same author.

In his 1941 dissertation (University of Minnesota) John W. Clark takes great pains to disprove the attractive "Pearl-poet" theory by examining the large quantities of data used by scholars from 1876 on, and finds their data invariably faulty or misapplied. Even granting them sufficient accuracy, Clark maintains their data could prove mere influence of one poet on another as well as it proves common authorship of the several poems.

But basically the fault in the early attempts at proving authorship lay in the inaccuracy of the data, which was necessarily gathered "by hand." "Precision without accuracy" is the

downfall of precomputer statistical analyses of literature. We should expect a computer-aided project to avoid this pitfall. One fairly typical study, Who Was Junius?, by Alvar Ellegard, does involve a computer, but the computer is put to work on data derived laboriously, intuitionally, and inaccurately by Ellegard's own hand. Therefore, his method, which purports to prove that the Letters of Junius were written by Sir Philip Francis, contains the traditional flaw of precomputer studies of the same kind.

The task set in the present project was to seek by objective means a method of proving authorship statistically by examination of vocabulary. The data were derived only from the known poems of five nineteenth-century writers, with no goal of attributing an anonymous or doubtful poem to any of the five. The purpose was to test only the test itself. "Precision without accuracy" was avoided by having the computer do the first selecting of vocabulary items to be subjected to analysis.

Fed into the 3600 Control Data computer were more than five-hundred pounds of IBM cards, each one bearing a single line of verse carefully verified, and with the spelling of certain potentially ambiguous words conventionalized in order to dispel the ambiguity: MAY=the auxiliary verb; MAYE=the month. A glossary program written by James D. Clark of Michigan State University gave me alphabetized lists for every text fed into the computer. A search of the largest texts having disclosed not one content-word significantly present in one author and not as significantly present in at least one other, I determined to use the common words in my testing.

Accordingly, I took the forty-five most common words and word-groups (AM+ARE+IS+WAS+WERE=one word; to which I apply the term allomorphs in my thesis) with the exception of the personal pronouns, and made of them a forty-five point profile for each of my 230 texts. If a writer subconsciously chooses one function word in preference to another, the one he chooses will form a peak on the profile-chart of his texts, while the other, less-preferred, word will show up as a valley. When charts of texts even of different word-populations are compared, two by the same author should have more points in common than two charts of texts by different authors.

The comparing was done by enlarging the points on the 230 profile-charts to quarter-inch holes, and then laying one chart on top of another and seeing where the holes coincided. Thus every text was compared with every other text, and five indices of correspondence tabulated for each comparison. A chart that finally analyzed the half-million bits of information thus tabulated showed that the method has possibilities: the best of four criteria into which I combined my five indices of correspondence was able to call the correct author sixteen times more frequently than would chance guessing.

The test might be highly reliable in establishing which author of only two wrote a doubtful text. Its value, however, for attributing a truly anonymous selection to one of a larger number of possible authors I will insist upon denying. To do this, a statistical test must have the same accuracy as a chemical test; it must work every time under laboratory conditions if it is to be

George W. Zimmer

assumed workable when conditions are less controlled.

Although the project was a failure, and the test is untrustworthy as it now stands, there is some possibility that by sharpening my procedures (perhaps by using the computer at all stages) and by strengthening the word-list by dropping dead items and adding some overlooked before, a reliable test for proving authorship could yet be discovered.

THE ATTRIBUTION OF AUTHORSHIP:  
A COMPUTERIZED METHOD EVALUATED AND COMPARED  
WITH OTHER METHODS PAST AND FUTURE

By

George W. Zimmer

A THESIS

Submitted to  
Michigan State University  
in partial fulfillment of the requirements  
for the degree of

DOCTOR OF PHILOSOPHY

Department of English

1968



3-18 67

Copyright by  
GEORGE W. ZIMMER  
1968

THE ATTRIBUTION OF AUTHORSHIP:  
A COMPUTERIZED METHOD EVALUATED AND COMPARED  
WITH OTHER METHODS PAST AND FUTURE

INTRODUCTION

One of the most tempting fields of scholarship is that involving the determination of authorship. If the historical-biographical approach to literature has ever had validity (and it shall here be considered a self-evident truth that it did and does have validity) it is necessary to know with certainty which author wrote which works. Our very important stereotypes of the various artists are largely dependent upon their canons of works: subtract a Timon of Athens from his canon and the percept "Shakespeare" becomes something other than it was with Timon in the canon. The authorship problem does, then, have its willing solvers, whose methods, however, vary widely.

The simplest method of settling an item within or outside a canon is by edict. By declaration of a scholar or a school of scholars an item is placed within or outside a canon. The logic for such edicts is partially as follows: a work can be placed in a canon only by the best authority; I am the best authority; therefore this work is declared to be part of the canon. Pragmatically, at least, the edict method is one of the best, especially when the placement of the work is agreed upon by those scholars most interested in the subject. But the edict method gives rise to the taking of exception by rival scholars or rival

schools. Then the rule states that the burden of disproof rests upon the dissenter who cannot therefore use the edict method. It is at this point that statistics are brought to play on the question.

"Methods using statistics" is a category classifiable in several different ways: "successful" and non-successful; simple and intricate; pseudo and true; early, more recent and most recent, or pre-computer, early computer, recent computer. Generally, the early attempts use simple processes involving pseudo-statistics and declare themselves successful. More recent attempts are less certain about success, although, paradoxically, they use scientifically sounder procedure.

To be dealt with in this thesis are species of early and recent, simple and intricate methods. Each species will be given its just share of criticism including the most recent which, according to my own classification, is also the most intricate and non-successful.

Sections of the thesis will show the deficiencies of several types of statistical methods for proving authorship, with particular attention to (A) the old, pre-computer methods that relied on pseudo-statistics and simplistic comparisons and that were so very positive in their results, and to (B) the computer-based method that involved me in my task. In between there will be short looks at hybrid methods: computer-aided and intricately statistical, but "positive" in their conclusions.

This element of "positivity" cannot be treated too fully.

The efforts of the scholars in the field of attribution of authorship in the pre-computer age have already been characterized as simplistic and pseudo-statistical. In general the earlier the study the less scientific it is; but in about direct ratio to the lack of scientific rigor, results of such studies tend to be declared "positive." There are very human reasons for these scientifically baseless declarations. In most instances the "positive" results support a preconceived notion. Later, as scientific rigor stiffened somewhat, an occasional scholar would claim to have had his opinion changed by his research, yet results would be set forth as positive, even though negative evidence was present in nearly as great proportion as positive. There are the further human inclinations to put an end to a piece of research, and to fill an unreasonable demand for a "happy ending" however false that ending might be. Such endings, which demand "positive" results, are easy to sell to a readership unwilling to study the data or the techniques upon which the results are based.

My dissertation presents in three chapters three stages or episodes in the quest for certainty by scholars in the field of English. The first chapter starts in 1941 and refers back to the beginnings of a particular authorship problem, that of the so-called Pearl-poet. It can be taken as typical of authorship problems, in that almost any argument adduced for the commonly-held opinion has been respected, while one most carefully prepared attack on the attractive theory has been all but totally ignored. The first chapter

quotes heavily from that attack (John W. Clark's doctoral dissertation, University of Minnesota, 1941) in order to represent its tone of futility and dogged determination.

Also, in presenting Clark's dissertation at such great length, I am repaying a debt I had not been aware of owing until I recently returned to its voluminous pages and found there many of the ideas and attitudes I had thought I had formed independently.

My second chapter advances to the 1960's, with an occasional flashback to Clark and the history of the authorship problem. In it I examine two types of computer-aided projects having as their ends the determining of authorship, and find both of them lacking integrity. It is suggested in this chapter that the computer should be allowed to do all of the work; especially in the first stages of a project, human error should not be permitted to intrude.

The third chapter, a delineation of my own project, which is a study of vocabulary for proofs of authorship, should follow the rule of maximum computer use. But it does not follow my rule. The discrepancy is owing to the fact that this dissertation was written backwards. My decisions on how to proceed had been influenced by Clark and many other books and articles<sup>1</sup> which I deliberately put aside while assembling my own project.

---

<sup>1</sup> Eminent among the writings are "Eras in English Poetry," Josephine Miles, PMLA LXX 853-75, which from examination of syntax posits three eras for each century of verse; Statistical Study of Literary Vocabulary. George Udny Yule (Macmillan, 1944), where formulae are given for determining the authorship

The project was finished and the results were tabulated before the relationships among the various methods fully occurred to me. I had intended at the outset to demonstrate an impossible thesis: namely, that it is not possible to prove authorship by statistical examination of vocabulary. The project, however, proved the only thing that it was capable of proving: that the particular method used here is not capable of proving authorship. The positive knowledge of this negative fact coupled with comparison of the method here originated with previous methods whose authors evince far less positive knowledge of the shortcomings of their work has now led me to a new hope that the authorship problem may yet after all eventually be solved by statistical means.

The results of my research in se were negative and could properly have been presented in a few pages. Although I am as reluctant as anyone else to write at length about little, I have a purpose for so doing in the sections of this thesis outside Chapter III. The positive result I aim at is putting an end to misguided work in the field, for which there may have been an excuse a few years ago, but not now. Chapter III, however, is no longer than it has to be, in contrast

---

of passages of 10,000 words or more; Type-Token Mathematics: A Textbook of Mathematical Linguistics. Gustave Morand (Mouton S'Graveharc, 1960), which concentrates on vocabulary items (types) and their occurrences (tokens) as the means of determining authorship; and "Montaigne-Shakespeare and the Deadly Parallel," George C. Taylor, Philological Quarterly XXII (1943) 330-3, which lists pseudo-seriously seventy-five kinds of evidence that canon and influence scholars use, of which seventy-five "vocabulary" is number 51.

to the too-explicit works I criticize. In keeping with this conciseness is my determination in Chapter III not to repeat, or to excerpt from, my charts and tables (thereby forcing the reader to consult the full tables in their proper context each time reference is made to one) unless to show how such excerpting can be made to give the appearance of validity to otherwise inconclusive results.

The fourth chapter is a retrospect of the three stages of work in the field, pre-computer, early computer and recent, with a look at some trends that started after this thesis was begun. It will there be shown how the problem of attribution of authorship will probably be solved, when and if it is ever solved.

## CHAPTER I

### A CLASSIC CRITIQUE OF PRE-COMPUTER STATISTICS

In March of 1941 John Williams Clark submitted for his doctoral degree from the University of Minnesota his dissertation, The Authorship of Sir Gawain and the Green Knight, Pearl, Cleanness, Patience, and Erkenwald in the Light of the Vocabulary. This work attacks the labors of all the scholars who up to that time had attributed the five poems to fewer than five poets. The first scholars interested in the problem had taken their cue from the manuscript of four of the poems: since the first four have come down to us in one document, the tendency would be to attribute them to a single writer. This attribution had been disputed several times before Clark, but the important editors of the texts had persisted in the one-poet theory, in support of which various kinds of evidence were presented. Professor Clark went to great pains to destroy his predecessors' arguments, sometimes by accepting their evidence but re-interpreting it, and other times by adding further evidence to show that they had used distorted data. His 581 pages of hard-hitting argumentation are comprehensive and attack every important argument adduced for the common authorship of any two of the five poems. Clark is thorough, completing and correcting



other supposedly scholarly work. He gives the history of his problem:

The earliest considerable attempt at deciding the question of single or multiple authorship came in 1876, in Moritz Trautmann's doctoral dissertation, Ueber Verfasser und Entstehungszeit Einiger Alliterierender Gedichte des Altenglischen; and the opinion there expressed--that the four poems were written by a single poet--was given currency a few years later by Ten Brink, in his Geschichte der Englischen Literatur, the English translation of which in the 'eighties extended Ten Brink's reputation, and the respect in which his pronouncements were held, beyond the then somewhat restricted circle of English scholars who willingly read German. Since that time, as Menner says (ed. Cl [for Cleanness] p. xi, n.1), "practically all those who have made special investigations of, or edited any of these poems . . . , have accepted the opinion" that a single author wrote Gaw, Prl, Cl, and Pat. This opinion--perhaps we may say this pious opinion--was erected almost into a dogma by being spread by Professor (later Sir) Israel Gollancz (who has done more than any other scholar to create the reputation--if not, indeed, the identity--of "the Pearl-poet") upon the sacred page of the Cambridge History of English Literature, of which the first volume, containing the account of the four poems, appeared in 1907. Sir Israel had published his adherence to the doctrine of single authorship as early as 1891 (in his first edition of Prl), and to the end of his life continued to proclaim it; his most absolute affirmation of faith being perhaps one that appears in the preface of his edition of Pat (1913): "It is now generally accepted, in respect of the four poems, that all the evidences of dialect, vocabulary, art, feeling, and thought, conclusively point to identity of authorship . . ." [Italics mine. [Clark's]]

The opinion that Erk, also, was written by the author of the four poems (or of one or more of them) was first advanced somewhat later (by Carl Horstmann, in the editio princeps of Erk, in Altenglische Legenden (Neue Folge), 1881), and has never been so widely adopted or, in general, so confidently urged, as the opinion that the four poems are from a single hand.

. . . In this attribution of the authorship of Erk to "the Gawain-poet," Horstmann (has) been followed by most scholars, notably R. W. Chambers, in Essays and Studies by Members of the English Association, 19. 126, n.2, but even Chambers does not express himself with assurance, and there is still a considerable amount of more or less

half-hearted dissent. (Pages 3-4)

Clark sees his task as giving heart to the dissent. He disposes in turn of each of the proponents of single authorship, and of each of the theories based on dialect (sixty-three pages), prosody (twenty-one), interests, attitudes and opinions (eight), syntax and style (thirty-two), and parallel passages (forty-one pages). After 192 pages, he is ready for Part II, the examination of the vocabulary. The chief contention that Clark advances in Part I is that similarities or even identity of dialect or the other criteria "prove" only that the writer of one poem lived in the same area as the writer of another, or was influenced by him. The evidence, he claims, cannot conclusively show either separate or common authorship, but, if the evidence is to be admitted, it has more force in proving diverse authorship. Always, however, there is the strong desire on the part of the earlier scholars to demonstrate the more attractive theory: to set up a "Pearl-poet" whose canon of works, together with those undoubtedly lost, would make him a worthy contender for honors commonly reserved for Chaucer.

Part II of Clark's dissertation is aimed at those scholars who, following Ten Brink's Geschichte der Englischen Literatur<sup>1</sup> and Sir Israel Gollancz in the 1907 Cambridge History of English Literature, sought to give a statistical

---

<sup>1</sup> History of English Literature: I, Eng. tr., H. M. Kennedy (New York, 1883).

foundation to their preconceptions. His strictures are particularly directed toward J. P. Oakden's Alliterative Poetry in Middle English<sup>2</sup> and Henry L. Savage, editor of St. Erkenwald,<sup>3</sup> although some lesser sinners attract a share of attention.

The vocabulary of the five poems early engaged the attention of scholars intent on discovering whether --or, as it almost seems as if we must sometimes say, on proving that--the poems were by a single author. Trautmann published on the subject three times. . . . Trautmann's conclusion was, in brief, that the vocabularies of the five poems are so much more like each other than they are like the vocabularies of any other ME alliterative poems, that we must suppose the five, and only the five, to have been by a single author. There is no doubt about the exceptional degree to which the vocabularies of the poems resemble each other; but subsequent investigations both more extensive and more exhaustive than Trautmann pretended to have made, or, indeed, could have made in the absence of editions with more or less complete glossaries and of NED, have shown that the vocabularies of the five poems are nowhere nearly so similar or so peculiar as Trautmann thought. This is well shown by Savage, in his edition of Erk, pp. liv-lv: ". . . the value of Trautmann's findings has been somewhat reduced by the appearance of the later volumes of the NED and the progress of scholarship; yet," Savage adds cheerfully, "the test of vocabulary indicates an unusually close connection between the five poems, and has strong affirmative bearing on the possibility of common authorship."

In other words, we are right back where we started --the vocabularies of the five poems are rather strikingly similar, but not by any means so similar that common authorship is the only possible (or even, I may add, the most probable) explanation. That this fact is perceived by the advocates of the theory of common authorship needs no further proof than that they all pay their respects to the vocabulary, and then look else-

---

2 University of Manchester Publications, CCV and CCKXXVI (Manchester, 1930 and 1935).

3 New Haven, 1926.

where for cogent arguments in favor of their view.  
(Pages 193-4)

And, as Clark demonstrated in Part I, the arguments from elsewhere are not cogent, either. Why, then, does he devote the bulk of his dissertation to an examination of the vocabulary, especially when "attempts to prove the common authorship of the five poems, on the basis of their vocabularies . . . have clearly failed"? (Page 195) At this point, Clark could take credit for examining the vocabulary, the most objective of all the criteria and the one at the same time that provides the most massive data; but instead he uses the device as a sort of tail to pin on one of his less preferred predecessors:

that failure has been proclaimed by no one more emphatically than by some of the principal advocates of the theory of common authorship themselves. But an opening has been left for studies of certain special aspects of the vocabularies; and that opening has been seen by the indefatigable Mr Oakden, who, in the second volume of his Allit. Poetry in Middle English, investigated three of these special aspects: (1) "Chiefly alliterative" words (by which Oakden means, not words that usually alliterate, but words "found but rarely or even not at all, outside the alliterative poems"); (2) "synonyms for man, knight," and (3) "synonyms to express movement." . . . consideration of his findings will repay our efforts by the further suspicions it will arouse as to the validity of the theory of the common authorship of the five poems, and will serve as an appropriate introduction to the main part of this dissertation. (Pages 195-6)

Clark soon shows that Oakden's fifty-three "chiefly alliterative" words reduce to but twelve that are thinly distributed among the five poems, only one appearing in all five poems; that at least one synonym (douth) demonstrates "a fundamental

difference in Sprachgefühl" (page 199) from poem to poem; and that Oakden's "definitely poetic" words for "go" have a characteristically non-significant distribution. Of some little significance for this present dissertation is Clark's explanation following his table of "go"-words for the five poems:

The glossaries [which Clark had thought he could rely on for his examination of Oakden's assertions] are complete except where "frequent" occurs under Gaw, and possibly also under Pat generally--here, as with the synonyms of Man, knicht, I have pieced out Bateson with Gollancz, and can hope for getting nothing more than an approach to the truth. All the zeros under Pat, however, are probably right; neither Gollancz nor Bateson, apparently, leaves words completely unnoticed in his glossary except by accident. (Page 213)

In this instance, Clark does not permit himself the dudgeon to which he ascends when he attacks Oakden directly; instead, he merely accepts the incomplete research for what it is not worth. Later, we see him re-doing far longer lists of French and Norse words, with the goal of achieving more nearly perfect accuracy. With an impossible foresight, he and all the other glossary-makers would have waited to let computers do the work with complete accuracy.

Yet, even accepting the imperfect lists for what they are, there must also be an accurate, honest reading of them. In one note, Clark accuses Oakden of carelessness (page 222), and of misreading "the glossaries, or silently 'correcting' them." The same note (no. 17, page 223) tabulates the etymologies given for four instances of the word note:

line	Gollancz	Savage	Oakden
38	OE	OE	OF
101	OE	OF	OF
133	OF	OF	OE
152	OE	OF	OE

. . . all three scholars never agree. It makes no immediate difference who is right, or whether anybody is right; the point is that neither Oakden nor anyone else really knows (and probably the author or authors of the poems did not know) how often this allegedly "chiefly alliterative" word from OE notu appears in the Five Poems. . . .

. . . As I have shown above, the editors of the poems sometimes disagree on the derivation of words . . . and yet neither Gollancz nor Savage (nor Oakden, for that matter) expresses the slightest doubt that the truth is attainable and that he has attained it. This sort of thing is common in the editions of the Five Poems. (Pages 223,4)

The slogan is "precision without accuracy," and the further point that Clark makes, about even the author not being aware of etymologies, is one to store away for future reference. However, he himself acts on the counter theory that the author(s) had an awareness of words, since the bulk of Clark's research is precisely in his lists of French and Norse words. Furthermore, Clark makes no claims to perfect accuracy in his lists. This disclaimer is in accord with the footnote above, but not with his abjuring of the "precision without accuracy" slogan.

It is ever thus. If the size of a writer's data is impressive, the statistics and whatever he does with them cannot fail to impress those who have no inclination to test them.

Clark had the inclination, and he did test Oakden's figures on the incidence of Old Norse words in the Gawain.

Let us consider Oakden's statement . . . that Gaw contains 236 ON words. What is an ON word? For that matter, what is "a" word? Are aslich, a., and asly, adv., two words or one, for our purposes? Much might be argued in favor of either answer; and so long as we give the same answer consistently, it makes no difference which we choose. But which has Oakden chosen? He doesn't say, and I don't know how to find out. Again, whichever he has adopted, does he stick to it? Again, I don't know. What I do know is that Oakden's habits of work, so far as I have observed them--and I have observed them pretty extensively--do not inspire me with confidence that he has given very careful consideration to the problem. (Page 353)

. . . I do not claim . . . that my judgment has been infallible, or even that it has been better than the judgment of Mr Oakden and of the editors of the several poems. But I believe that I have shown, beyond reasonable doubt, that Oakden (like most of the editors) has overestimated the number of words, in the five poems, of which we can say with confidence that they are probably ON. (Page 355)

The aspersion cast (page 353), Mr. Clark concludes:

Incidentally, the close similarity of Oakden's figures and mine for each of the poems severally (except Gaw) leads me to believe that Oakden must, after all, have been nearly as cautious (always with the exception of Gaw) as I about calling a word ON; (page 357)

thus he puts himself in the same pocket as his chief target:

such statements as Oakden's that Gaw contains 238 ON words--or mine that it contains 202--are perfect examples of precision without accuracy, the fact being, of course, that no one knows how many "distinct words" the language contains, or how many of them are ON. (Page 358)

A weird sense of futility pervades the dissertation. The Norse words disposed of, in a mere 120 pages, Clark turns to the French words. His predecessor, Hartley Bateson,<sup>4</sup> had worked out the proportions of French words in the two poems

---

<sup>4</sup> Ed., Patience (2nd ed., London, 1918).

Pearl and Patience as 34.47 to 19.92.

Bateson, like Ogden with ON, either is being precise without being accurate, or is neglecting to give notice of his realization of that fact; but since the words probably miscalled OF are, compared with those probably miscalled ON, very few (fewer than 30), and since there are several (probably about five) times as many OF words in the Five Poems as ON, Bateson's ratio of the OF element in Prl to the ON element in Pat may be taken as reasonably close to the true one--as close, probably, as any that could be arrived at. (Page 468)

Then, since he is unsure of the meaning of Bateson's ratio of 34.47 to 19.92, Clark does it over again; "and for good measure, I have extended the investigation to the other poems." (Page 469) He laments:

A dissertation, I suppose, is no place for lamentations about the exasperating tediousness of what was, after all, a self-imposed task; but I cannot forbear to say that I wish (a) that editors of ME texts had agreed on such questions as the proper alphabetical position of ȝ (in both values), y (vowel), and u and y, and (b) that ME poets and scribes had, as Artemus Ward said, known how to spell. The mere mechanical complications introduced, by these two factors, into the compilation of such a list as that below, are beyond the powers of anyone who has not experienced them to imagine. But the job was at length finished; and I have the comfort of knowing that I have discovered some facts from it, and subsequent students--whose mathematical studies were not cut short, like mine, with plane geometry--may discover more. (Page 469)

But what has he discovered that is of any lasting worth? His "statistics" are of no more value than those of the scholars he attacks: the data are gathered with almost as many opportunities for error as were present when the earlier scholars worked without the aid of the Oxford English Dictionary, and the data are handled in a most unstatistical manner. Clark works out ratios of French to Norse words for



each of the five poems. Pearl has 4.29 OF words to 1 ON word; Cleanness 3.86; Erkenwald 3.60; Gawain 3.56; and Patience 2.67 OF words to 1 ON word.

It is my duty to remind the reader once more that these figures, like all those in this chapter, are only approximate, so that such a difference as that between Erk and Gaw, or even that between Erk and Cl, is probably not very significant; but such differences as . . . that between Prl and Pat can certainly not be attributed to the roughness of the basic figures. (Pages 478-9)

Perhaps not, but reason for tabulating the ratios in the first place can be questioned, and the differences can hardly be said to indicate anything of importance. Clark sees the difference between Patience and Cleanness as showing that the former was written first, "before the author had become so well acquainted with French literature," (page 479) although an increased sophistication about French writings could conceivably have caused an author to eschew French terms in favor of native (including, all unaware, Norse) words.<sup>5</sup>

Again, Clark is precise though inaccurate when determining numbers of French words, counting as a single form the noun and verb taken from the same source-word, but counting them as two if they are derived independently (affray, n. + affraye, v. = one word; but afyaunce, n. + affye, v. = two words). (Page 481) Furthermore, he counts separately the full and aphetic forms of the same word: "This is perhaps not entirely reasonable, but it is convenient, and can hardly make

---

<sup>5</sup> Later (page 555), Clark admits that it is "pointless and unrealistic for us to pretend to know when a ME poet was thinking of the native word and when he was thinking of the foreign one."

any serious difference in any conclusions to be drawn from the list," (page 480) the reason being, that Clark himself does not take his lists very seriously. When you are out to disprove a theory, any kind of statistics will serve. Nevertheless, I am happy that he did distinguish the full and aphetic forms, because thereby he "proved" the greater proportionate incidence of the full form in rimed poetry. His care with the two kinds of Middle English poetry no doubt influenced my decision to consider verse forms in deriving glossaries. Just as Clark wasted effort in verifying Bate-son's ratios, so also did my meticulousness go for almost naught. (See Chapter III.)

I find, upon re-reading Clark's dissertation four years after I first borrowed it (during which four years my own project received its form), that I am more indebted to it than I would ever have cared to admit. The impression I retained was almost entirely of his contentious tone stemming from exasperation and frustration. I see now that it was he who must have put into my mind the notion to settle upon the inconspicuous words as possible indicators of writing habits. From him came my technique of so dividing poems that the effect of rime could be either emphasized or nullified. And, finally, Clark's dissertation had, as I hope mine will have,

the value of pointing the way to a method and a point of view, in determining questions of authorship, that may, when brought to perfection, lead to more positive and dependable conclusions than I have been able to make them yield. (Page 572)

His method, owing to his negative point of view, yielded Clark further frustrations. Positive thinking has a far better chance at publication than does mere objectivity. Or perhaps Clark's tone insulted too many too important scholars.<sup>6</sup> The dissertation was never published in its entirety. Portions of it appeared in a variety of journals,<sup>7</sup> but his powerful argument was completely ignored fifteen years later by a supposedly demolished target when Henry L.

---

6 His targets include Gollancz, Savage, Oakden (of course), Miss Mary Serjeantson ("The Dialects of the West Midlands in Middle English," in Review of English Studies 3. 54, 186, 319 (1927)), J. R. R. Tolkien and E. V. Gordon edd., Sir Gawain and the Green Knight (London, 1925), and Lane Cooper and his doctoral students. Oakden and Serjeantson are accused of a suppressio veri (page 35), as are Tolkien and Gordon (pages 414-5). Oakden is treated quite harshly throughout, often through the medium of jokes and catch-words, the points of which are not always clear to me. Perhaps most typical of Clark's fits of ill-humor is the footnote on page 74: "This is neither Mr Chapman's first, nor his most ambitious, nor his most fruitless contribution to the study of the four poems. The only true and lawful claimant to those titles is Mr Chapman's doctoral dissertation, produced in 1927 at Cornell University under the Concordantifex Maximus, Professor Lane Cooper, and entitled 'A Lexical Concordance of the Middle English Pearl, Cleanness, Patience, and Sir Cawayn and the Grene Knight.'" In a five-page preface (the only thing the dissertation contains besides the concordance proper), Mr Chapman writes: 'The work is recorded on about 40,000 slips ...; and in this shape the copy in due time will be sent to the printer...' Clark is piqued because this concordance, which he intended to rely upon, covers only six letters of the alphabet. Nowadays they use the computer at Cornell for the production of concordances. A quality product results, but without that useful byproduct, the Ph.D.

7 "Observations on Certain Differences in Vocabulary Between Cleanness and Sir Gawain and the Green Knight," Philological Quarterly, XXVIII (1949), 261-73; "Paraphrases for 'God' in the Poems Attributed to 'The Gawain-Poet,'" Modern Language Notes, LKV (1950), 232-6; "'The Gawain poet' and the Substantival Adjective," Journal of English and Germanic Philology, XLIX (1950), 60-6; "On Certain 'Alliterative' and 'Poetic' Words in the Poems Attributed to 'The

Savage<sup>8</sup> again advanced the theory of common authorship as being all but universally accepted.

---

Gawain-Poet, " Modern Language Quarterly, XII (1951), 387-98.

8 Henry Lyttleton Savage, The Gawain-Poet: Studies in his Personality and Background (Chapel Hill, 1956). John Conley's review points out that "John W. Clark's sobering studies of the vocabulary of the five poems are not even listed. Yet Professor Clark shows that J. P. Oakden's Alliterative Poetry in Middle English, to which Dr. Savage appeals, is far from being trustworthy." Speculum, XXXII (1957), 858-61.

## CHAPTER II

### AUTHORSHIP PROBLEMS IN THE EARLY COMPUTER AGE

The main quarrel that a statistician would have with Clark or his predecessors would be over the tiny samples that they sometimes dealt with. Secondly, the statistician might marvel over their lack of an appropriate formula that takes into consideration populations and deviations. Finally, he could ask for a kind of accuracy that neither Clark nor his predecessors were prepared to provide. Pre-World War II scholars could not have foreseen the computer, and put off their researches until it became available. But the authorship question continued to interest scholars even into the nineteen-fifties and -sixties, when a number of studies were made which used the computer for some stage or other of the operation.

Who Was Junius? and its companion volume, A Statistical Method for Determining Authorship,<sup>1</sup> by Alvar Ellegard (statistics by Per Sigurd Agrell), both make heavy use of the computer. The latter volume is almost wholly concerned with tables of statistics worked out by computer, with accompanying explanations. The pair of books is a convincing perfor-

---

<sup>1</sup> Stockholm: Almqvist and Wiksell, 1962.

mance. Ellegard's history of the problem shows that Philip Francis (1740-1813) had always been the prime candidate for the honor of having composed the Junius letters, although Ellegard claims that he began his study favoring another possible writer. Ellegard's central point seems to be that direct testimony by the contemporaries of Junius can, now that those witnesses are dead, have no chance of pointing out the writer to blame for these incendiary pamphlets, and that therefore a more objectively deductive method must be used to determine their authorship.

Sir Philip Francis had long been the favorite candidate among scholars who conjectured an author for the Letters. Ellegard was persuaded by the biographical and other evidence that the chances were against his having been the author, and he set out to prove his contention by statistical means. His method has several faint echoes of techniques that Clark examined in his critique of the scholarship on the "Pearl=poet." Clark's unpatentable concepts of the favorite word<sup>2</sup> and of the unconsciously-chosen expression<sup>3</sup> alluded to at the end of the previous chapter form the base of Ellegard's system.

Simply, his system would compare corpora of Junian writings with non-Junian writings by the presence or absence of certain words or expressions. These plus- and minus-words -- originally 458 of them, later reduced to 272 -- were

---

2 Clark dissertation, page 302.

3 Pages 9-11.

culled rather subjectively<sup>4</sup> from all of the Letters, from all of the known, identifiable writings of Sir Philip Francis, and from a "million-word sample" of contemporary writings. A Junian plus-word is one used with higher frequency in the Letters than in the "million-word sample" and a minus-word is one that occurs with less frequency in Junius than in the sample. The 458 words are on a sliding scale of positiveness or negativity, with words used not at all (or almost not at all) in the Junian letters designated the most minus. Ellegard himself culled the tables of occurrences of these 458 words or expressions from double readings of all of the texts involved (Junius, Francis, the million-word sample). From these tables came the raw data which were fed to the computer for manipulation by formula and for multiple grouping.

To repeat, the books by Ellegard and his statistician are a convincing performance. I wrote a too-favorable review of them for the Journal of English and Germanic Philology,<sup>5</sup> asserting my belief that Ellegard had solved for all time the problem of the authorship of the Letters of Junius. I dampened my praise, it is true, by mentioning a factual error or two, and by not accepting completely either Ellegard's objectivity or his ability to scan so many texts for his 458 items as accurately as he would have his readers believe he did. My review was in the mail to the Journal of English and

---

<sup>4</sup> Despite Ellegard's claim of objectivity; see below, pages 26ff.

<sup>5</sup> June, 1963, pages 688-9.

Germanic Philology before I saw the handling of the books by the reviewer for the Times Literary Supplement,<sup>6</sup> who brought a nicer skepticism to the task.

I was then working on a series of projects for Professor Arthur Sherbo at Michigan State University, taking over in the midst of one project that had already been started with another assistant. The goal of these investigations, as I understood it, was to determine which of several 1000-word texts belonged to Samuel Johnson and which were spurious.<sup>7</sup> Therefore my interest in Ellegard's apparent success was colored by my hope for a like success, and slanted by my involvement in a Johnsonian project.

Looking at the Ellegard books now that I have become more blasé in the face of ponderous scholarship and somewhat conversant with, if not statistics, statisticians, I can well believe that Ellegard's statistician was sneering at him, (he is called "subtle" by the Times Literary Supplement reviewer) while working up his tables of results. For statisticians seldom put any credence in the statistics of laymen, particularly when those laymen are scholars of literature. One noted statistician with whom I talked at Michigan State University even refused to accept the widely accepted results of

---

6 "The Statistics of Style," January 5, 1963, page 1.

7 Though the sampling was wide enough for an effective analysis, having twenty-two texts (mostly Johnson's, but with examples from the known writings of other possible authors of the disputed texts), it was not deep enough: a thousand words could not provide sufficient data for the relatively crude tests to which they were to be put. There will be more about these projects later.



the work done on the Federalist papers by Frederick Mosteller and David L. Wallace,<sup>8</sup> where the problem involves only two possible writers.

The professional statistician's skepticism stems from the tendency of dealers in authorship problems to want to sell their cause regardless of negative evidence. Thus, while Ellegard claims to have been converted from an anti-Franciscan stance, nevertheless his case is made to seem inevitably to lead to Francis. And although he may have chosen his plus and minus test words from the Letters before he believed them to be by Francis, he still knew Francis as a prime candidate, and he could have been subconsciously influenced in his choice of the 458 test words by his familiarity with Francis' writings.

I can cite no authority for the caveats in the preceding paragraph. The notions expressed formed in my mind as my belief about Ellegard and others changed under the influence of one or both of the statisticians to whom I talked at Michigan State University. Some statisticians -- and again I cannot cite texts or give names -- would further object to the way the analysis in Ellegard's books groups writers and texts. The million-word sample contains texts which are then regrouped into the writings of individual authors, and compared with the million-word sample. There is a bias factor in the formula used to evaluate the deviations of individual authors'

---

<sup>8</sup> Inference and Disputed Authorship: The Federalist, (Addison-Wesley, Reading, Mass., 1964).

writings from the standard of the sample. Naturally, Francis is found to deviate farthest from the sample and to be the closest to the Junian standard.<sup>9</sup> To one of the statisticians with whom I spoke, a text is a text; and it loses its testability when divided or when combined with other texts. Ellegard's system repeatedly combines texts.

The 272 test-words are themselves combined also. The changes are rung on the combinations of groups of groups of plus- and minus-words. It is not enough that there is a descending scale of Junianity attaching to groups of plus-words; in order to test many texts in a variety of ways -- he fills forty-five pages with tables -- Ellegard combines his groups into the very Junian, the somewhat Junian, and the slightly Junian. Of course, the first super-group excludes non-Junian texts better than the other two. That this is a species of circular argument is not pertinent, since the results, by such manipulation, are rendered so much more positive.

Putting aside the subtleties read between the lines by the Times Literary Supplement reviewer of the Appendix written by Ellegard's statistician, Per Sigurd Agrell, perhaps the main mark against the theory, even assuming that it proves Francis the most likely author among those authors considered, is the contention that only a few of the several

---

<sup>9</sup> The possibility that Francis might have consciously imitated Junius is not given weight.

hundred thousand Englishmen capable of having written the Letters were entered into Ellegard's process. Ellegard here rightly falls back on biographical details. Francis was available and knowledgeable as the letters were being written, and they stopped when he went to India. Yet the fact remains that almost any number of secretaries or mistresses of Opposition members in Parliament could have had access to the information displayed, and might have had the literary skills associated with Junius.

But if the "'one new fact' demanded by Dilke"<sup>10</sup> is provided by Ellegard's finding that Francis' style is the closest to the Junian of all the feasible and known contenders, perhaps the question is after all solved. The method, however, is not convertible to other problems of contested authorship where the biographical details are not so clear-cut. It would be almost worthless where the data-providing subject matter is limited, where texts contain fewer than two thousand words, or where the writings are in verse.

My principal criticism of Ellegard's procedure concerns his method of gathering data.

My procedure was as follows. After a cursory reading of the whole material -- all the Junius Letters --, in order to get a general impression of the language of the time, I carefully combed the Junius material for words and constructions which seemed to me to be used with remarkable frequency in it. I then did the same for the comparative sample of a million words, noting not only the words which struck me as remarkably frequent in the various texts, but also those words which,

---

<sup>10</sup> Who Was Junius?, page 119. Charles Wentworth Dilke wrote in the Athenaeum in the mid-1800's of his disbelief in the theory of Franciscan authorship of the Letters.

though not particularly frequent, I did not remember having seen in Junius.

In this way I obtained a preliminary list of Junian plus words -- from my reading of Junius -- and of Junian minus words -- from my reading of the comparative material. After I had got the whole list of plus and minus words by heart, I read through the whole text material again, registering each occurrence of each word included in the list. When this had been done, the total number of occurrences was added up for each word, in Junius on one hand, and in the comparative material on the other. After this, it only remained to calculate the distinctiveness ratio, and the final testing list could be drawn up.

. . . . .

In order to minimize as far as possible the number of occurrences lost by inadvertence, each page of text was read through twice over. Even so, however, mistakes have certainly been made.

At the very time I was reading Ellegard's book for review, I was having trouble maintaining accurate counts of my own. The old MISTIC computer seemed not to be as truthful as we thought it should be, in giving us counts of words in sentences and of words of certain lengths. I wondered if the program for alphabetized lists of words was also playing us false. The text I checked was supposed to have thirty-four instances of the word "of" according to the count of the MISTIC computer. I then discovered how difficult it is to find as few as thirty-four "of's" in a text of a mere thousand words; and I knew the number that I was trying to find, and was circling each "of" as I found it. It took about eight readings. Ellegard's assurance that two readings would not result in mistakes that would have significant bearing

on his results<sup>12</sup> is one statement that I cannot accept.

Ellegard was aware of the means available for avoiding all mistakes, in deriving an unbiased testing list, as well as in counting the items from that list in the separate texts.

There are two ways of guaranteeing an unbiased testing list. One is to examine completely the vocabulary of all the texts investigated, and draw up the lists of plus-expressions and minus-expressions wholly on the basis of such complete investigation. The other way is to select a sample of expressions according to a well-defined objective criterion, which can be reasonably assumed not to favour or disfavour any particular candidate. The selection may be either random, or systematic.

I have followed neither course. To make a complete investigation would have been a Herculean task: it will have to wait until the whole process can be carried out by electronic computer. . . .<sup>13</sup>

Ellegard's choice was wrong. Unlike John W. Clark in 1941, he could have chosen the computer and he should have waited for suitable programs for sifting and culling. Instead, he repeats the error that Clark is so vociferous about of having precision without accuracy. His indices to the fifth and sixth decimal place, and his forty-five pages of perfectly worked-out tables are all based on shifty data. The precision and accuracy must begin at the beginning or there will obtain inexorably the dictum of the computer operators: Garbage in, garbage out.

Computer-use for authorship problems was very much in

---

12 A Statistical Method . . ., page 23.

13 Who Was Junius?, page 113.

the atmosphere in the early sixties. The first phase of the Johnson problem initiated by Professor Sherbo ground to a halt: the thousand-word texts we were using were just not large enough to show distinctiveness in such aspects as sentence-length, or numbers of x-lettered words. We were looking for a determination of which of several disputed texts were Johnson's by means of a statistical process suggested by George Udny Yule.<sup>14</sup> Even the excellent glossary program which gave us the occurrences of every vocabulary item in alphabetized lists failed to yield promising results, presumably for the same reason: the shortness of the texts. We then struck out in a new direction. Retaining the twenty-two thousand-word texts while finding some way to increase the kinds of possibly distinctive data was the immediate task. White Knight that I am, I invented my own system.

The system sought to multiply the data by counting, not words, but groups of words. The repetition of patterns might be the clue to an author's writing habits. Raw words would, of course, not serve this new purpose, because there is so much variety in the selection of content-words that the longer patterns would almost never be repeated. The raw words were given coded designations indicating their "part of speech" and their "function in sentence." Thus the

---

<sup>14</sup> A Statistical Study of Literary Vocabulary (Cambridge, Macmillan, 1944), and "On Sentence-length as a Statistical Characteristic of Style in Prose; with Application to Two Cases of Disputed Authorship," Biometrika, XXX (1938), pages 363-90.

expressions "in the house," "on the town," and "over the rainbow" would all have the same coded appearance: PJ DE LP, for "Preposition introducing adJective phrase," "the Determiner 'thE,'" and "Noun object of Preposition." All 22,000 words had to be so coded (by hand!) before the computer program for selecting patterns could begin to work. By this time the casual reader will recognize the same old fallacy of precision without accuracy in the work. The coding had to be done by hand, and I made the usual claim of consistency to refute in advance any argument that my data might be deficient.

The program ran, the like patterns were collected, and the output was analyzed. The results were nothing if not inconclusive.

Not so, however, with the results of a very similar investigation that was being carried on at the same time at Columbia University by Louis T. Milic. Mr. Milic's dissertation also examined prose of the eighteenth century by tabulating patterns of words;<sup>15</sup> but his coding of words was based on the structural grammar of C. C. Fries rather than on the traditional parts of speech, and his sample texts ran to four-thousand words instead of our one thousand. Perhaps, I reasoned, his results would tend to be more positive than ours.

---

<sup>15</sup> A Quantitative Approach to the Style of Jonathan Swift (The Hague, Mouton, 1967). Mr. Milic and I exchanged several letters and talked long distance once as he was putting the finishing touches on his dissertation. I can at this moment understand his state of mind the day my call went through to him. By the time his work would have been available, I had already abandoned the system.

Despite Milic's apparent success, (his "study claims to have produced a method of identification by internal evidence, free of the usual uncertainties, using statistical methods and computer technology")<sup>16</sup> I decided to reject the process we had discovered independently when it came time to undertake my own dissertation project. What course I took, how it failed, and why I think yet that it is, in the main, the right course form the substance of the remainder of this thesis.

---

16 Dissertation Abstracts, XXIV (1964) 3730.



### CHAPTER III

#### DISCOVERING A TEST FOR PROVING AUTHORSHIP: A STATISTICAL TREATMENT OF MOST OF THE LONG POEMS OF WORDSWORTH, KEATS, SHELLEY, BROWNING, AND TENNYSON

The project I now undertook started out to be the impossible one of proving that authorship cannot be proved by a statistical examination of vocabulary. Translated into possibilities, it meant that I would devise the best test I could and use it under ideal conditions with the hope that it would work but with the expectation that it would not.

Of all the segments of the initial project with Professor Sherbo, I had confidence only in the "glossary" program. The tests for sentence- and word-length were almost patently unworkable when used on our short, thousand-word texts. With alphabetized word lists, however, texts could be compared for every item of vocabulary.

Such was my intention: to omit consideration of no word out of fear of the charge of conscious or unconscious prejudice that Ellegard was subject to when he drew up his list of 458 items. I also left the Johnsonian milieu with its doubtful texts so that I could concentrate on the test itself and not on any immediately practical application thereof.

To the best of my limited knowledge, there had been no completely objective tests of authorship. Even if the researcher

is without an axe to grind, so to speak; that is, even if he does not hold a prior belief that a certain author is to be credited with the disputed work, he nevertheless does begin with a strictly limited set of candidates, and omits from consideration the possible stray contributor, or the truly anonymous writer who was not known in his own time to have written anything. Always, in such research, there is the task set of attributing something of doubtful authorship to a known writer. And although controls are purportedly used, the methods are never tested entirely apart from the problem for which they were designed. Having a goal in mind can cause an experimenter to color his data, even unconsciously, by selecting items for analysis any other way but at random.

Another abuse of scientific methodology occurs when comparatively scant data, never gathered with perfect accuracy, are formulized and magnified into imposing tables of figures to the third and fourth decimal place. This could be called the Gold Bug distortion, whereby a mistake of inches near the trunk of the tree amounted to many feet when the final line was projected. The deeper you dig in a wrong location, the more foolish you appear in retrospect. Ellegard's two volumes on the Junius problem are a good example of this sort of abuse.

I take pains to avoid both of these pitfalls. My purpose is to seek a method of proving authorship by examination of vocabulary usage. I avoid the first trap by choosing to examine only known works by known poets. And, secondly, I allow

the perfectly accurate counting machine to amass my initial data, from which I subtract, by wholly objective means, the usable parts.

Can a poem by Keats be distinguished from one by Shelley, Wordsworth, Browning, or Tennyson through the use of a test involving the poets' choice of certain vocabulary items? With the aim of discovering such words, I set about to feed every poem of more than two thousand words by these five nineteenth-century poets into the 3600 Control Data computer then newly installed at Michigan State University. I did not have enough time to submit every poem of more than two thousand words by the five poets, but my omissions were completely by chance (see Appendix A for a list of the texts). If an authorship test by vocabulary does exist, it probably will not work on poems of much less than four thousand words, but my intentions included the determining of how small a text can be tested successfully.

Each line of each poem was punched on an IBM card and carefully verified (see Specimen 2 in Appendix F). Not knowing which words would enter into my analysis, I sought to eliminate ambiguity by following certain conventions of spelling. The auxiliaries "may" and "might" were to be separated from the nouns of the same spelling by appending an "c" to the nouns. The British "round" was always spelled "around" when it meant "around." Contractions were spelled out so that both parts of the word could be counted while the "word" itself would register only once: "itis" for "it's," and "cannt"

for "can't." My failure to somehow differentiate "to" the preposition and "to" the infinitive signal is only slightly mitigated by the contention that all "to's" are equal inasmuch as the poet is choosing in either instance the same two-letter word.

Mr. James D. Clark, of the Department of Psychology at Michigan State University, wrote the program for data retrieval. Mr. Clark's glossary program yielded me alphabetized lists of words and their occurrences in more than two hundred texts (see Specimen 3 in Appendix F). The possibilities for expanding the number of "texts" are nearly limitless, since halves, thirds, fourths, or sixths can be combined in many ways. No text is made up of parts from different poems, however.

With the computer output I was ready to follow Ellegard's retrospective advice to select words entirely objectively (see quote, pages 26-7). I conned the lists for words significantly present in texts by one poet and not so present in texts by the others. No such words seem to exist; that is to say, if a word is used in several poems by one writer, it will be used with about the same frequency by at least one other writer. It would be necessary, I decided, to work with those words appearing in practically every text. These are the words automatically excluded from most concordances: the non-content or function words. And since they all appear in almost every text, my treatment of them would have to consider their relations to one another: does a writer's repeated choice of "the"

diminish his uses of "a" while at the same time, perhaps, his "and's" are impinging upon his "or's"?

I had thirteen of my texts (see Appendices A and B) of about eight thousand words scanned for the words common to all thirteen. After the disqualification of personal pronouns as too dependent upon content,<sup>1</sup> forty-five words remained. Fourteen of the forty-five have variant forms, which were carefully combined to make single items (see Appendix C for the allomorphs of these fourteen items). It was not possible at this stage to distinguish the usages of several ambiguous forms, such as "as" and "like," but even deliberate refusal to distinguish them could be justified by the argument that the poet did after all choose the word in question, and probably unconsciously, since most of the forty-five items tend to take light stress in their verses. This justification could perhaps extend to the single item "to," which might have been separated into its use as preposition and use as infinitive signal.

The forty-five key words having been determined upon by purely objective means, it remained to find a way of using

---

<sup>1</sup> This decision is based on an experience that Professor Sherbo shared with me. An examination of three 12,000-word texts from consecutive issues of The Gentleman's Magazine of the 1740's for clues to identify the author of the doubtful middle text showed that the single word to vary significantly in usage from text to text was the feminine pronoun. One of the articles was about the Queen of Spain and also used the feminine pronoun for certain abstractions. Another article also used the feminine pronoun for abstractions, but was not concerned with the Queen. The third had no feminine pronouns, although some of the same abstractions were referred to.

them for comparing texts. It will be remembered (Appendix A: list of texts) that the texts vary in length from one thousand words to eight thousand. I imagined that there could conceivably be distinct patterns of usage for these test or key words. Such patterns would have to have been determined entirely by unconscious selection by the poets. If, or since, they were beyond the ability of the poet to control, these patterns should be so much the more effective for use in establishing an authorship test that would distinguish a poet from his imitators.<sup>2</sup> And if such a test existed, it could possibly be used on texts of vastly unequal length. So, rather than concentrate on comparing texts of commensurate size, I decided to compare each text with every other.

Accordingly, I made profiles of all the texts by graphing the forty-five words. I gave the word most frequently used in a text the value of 100%, plotted at the top of the graph. Each of the other forty-four words were given proportionate positions (see Specimen 6, Appendix F). In order to compare

---

2 My interest in the subject somewhat antedates my quick Masters Thesis (University of Detroit) written in the summer of 1959 in which I "traced" "evidence" of "influence" of Shelley on four subsequent poets, mainly through their use of common vocabulary items. The kind of item I then concentrated on is typified by the word skiey used by Shelley, of course, and by Francis Thompson in what must have been a conscious attempt to resemble Shelley. A word would not have to be as outrageously "poetic" as skiey for an admirer to borrow it; the other conscious borrowings would, however, tend to be the slightly out-of-the-ordinary. Apart from the function words in "turns of phrases" so borrowed, practically all of the borrowings would be content words. And no imitator would think to conform his own usage of all function words to the patterns of his master.

the graphs visually, I enlarged the forty-five points to quarter-inch holes (see Specimen 7, Appendix F). Properly positioned, one on top of the other, all the points of comparison between two graphs were immediately visible and ready to be counted. A gross count of simple coincidences would not have justified the use of graphs, since such data could be compiled merely by having the computer examine the charts of numbers that lay behind the graphs, and letting it do the ratios at the same time. Perhaps, I reasoned, the best profile similarities were skewed out of recognition by the lack of coincidence between the two leading words, that is, those given the value of 100%. Frequently it did happen that the greatest numbers of closest correspondences were found only after searching for them.

This searching added but little time to the comparison of each pair of graphs. Each comparison of two graphs took approximately one minute. After taping Graph #1 to a dark board, I positioned Graph #2 atop it and counted (1) the number of holes that corresponded at all, (2) the total that corresponded closely (that is, that showed more than half a hole-diameter), (3) the number of correspondences above an arbitrary 15% line. For the fourth and fifth indices of correspondence, Graph #2 was slid up an inch toward the top of the board while I looked for the greatest number of additional correspondences above the 15% line on Graph #1. Then Graph #2 was slid down an inch (two inches, really) while correspondences above the 15% line on Graph #2 were sought.

Sometimes more than a minute was consumed in finding the scores for the fourth and fifth indices -- approximate and close correspondences above the 15% line. Time was also consumed in taping the bottom graph to the board and removing it, in taking out and putting away the sets of graphs and the sheets onto which I was writing the five indices. I wrote five index numbers (sometimes seven) for each of  $230 \times 229 \times \frac{1}{2}$  comparisons of graphs, and I counted about fifty correspondences for each comparison of two graphs. The 1,320,000 bits of information thus counted were recorded on a triangular chart made up of three hundred individual  $8\frac{1}{2}$ " by 12" sheets, and measuring twenty by twelve feet (see Specimen 8, Appendix F: a part of one of the  $8\frac{1}{2}$  by 12 inch sheets).

A computer, which would not have had to tape graphs to a board, or would not even have had to use graphs at all, could have completed the counting in a matter of minutes once it had been programed and the material had been prepared for it. I justify my performing this long phase without the aid of a computer by the fact that I was not exactly certain what I was looking for or how I would be able to use the data I was compiling. At one stage of the comparison phase, when it was about one-fourth finished, I took notes on how a certain text (#36 by Shelley; see Appendix A) compared positively with other texts: that is, with what texts did it yield high indices of correspondence. I was most exhilarated when, seven



times out of seven, high indices actually did point to other texts also by Shelley. But the eighth, tenth, and eleventh comparisons of the twelve I made resulted in false identifications. So promising did the system appear at this time that I attempted to present an explanation of it at the April 30, 1966 meeting of the Midwest Modern Language Association in Iowa, where I learned that it is nearly impossible to present unconvincing facts convincingly. For by the time of the conference, when my facts should have been more firmly positive, because I was by then dealing with larger texts, they were more inconclusive than ever before or since: I knew they were not as positive as the Shelley #36 figures indicated, but neither could I say that they would be negative until I had finished the three hundred pages of my 20' x 12' chart.

With all the indices tabulated, the final step was to test the results. If there is a profile or a set of profiles made up of points on a graph representing the proportional occurrences of forty-five common words selected objectively, and typical for each of my five poets, then surely the following test will find it. I reduced the five indices for each comparison down to four criteria, three positive and one negative. The first positive criterion consists of the five indices added together, the second is merely the third index (correspondences above the 15% line), the third adds the last two indices (the moveable ones), and the last is the same as the first, except that the low total is the test. That is to say, if two graphs have no points of correspondence then the

poems behind the graphs should not be by the same poet. Any of the positive criteria should identify texts as being by the same poet, since what I looked for in each case were the extreme examples (combination of indices distorts the data in my favor; it was done because there was no way of predicting which criteria or indices would assay out). For each of the 230 texts, I rejected all but those texts of the remaining 229 that were most like it. By expectation, at least above some limit of about 3,500 words, each text should have selected matching texts; each Keats text should have selected Keats texts, each Shelley, other Shelleys. Beyond this, the negative criterion should never have selected texts by the same poet.

The chart of the last analysis (see Appendix E for a summation of this chart), showing the results of one of the grossest possible tests of the validity of my method, must measure 230 by 230 squares. Most of these squares will not contain an entry because only the extreme examples of merely four criteria are tabulated for each row of 230 squares. Let me present here (in Figure 1, next page) what might be a random sample of the chart of the last analysis. Divided into a hundred equal parts of twenty-three squares on a side, one such part of the chart, the twenty-third (counting from top left), contains twenty-one criteria of correspondence, every one of which indicates texts that are indeed by the same poet, or, in the case of the negative criterion, texts not by the same poet. These are exactly the results I had hoped for. The next step

	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69
	W	W	W	W	S	W	T	B	W	S	W	W	W	W	W	W	K	W	W	W	W	W	K
47W													2										
48W																							
49W	II											3II											
50W																							
51S																							
52W	3			2																			
53T												0											
54B	0	0				0													0	0			
55W																							
56S																II							
57W									3				3										
58W			3II																				
59W	2																						
60W																							
61W																							
62W																							
63K																				3			
64W																						0	
65W																							
66W																							
67W																							
68W	2																						
69K																							

# KEY TO CRITERIA

- 2 = total of 5 indices
- 3 = 3rd index: correspondences above the 15% line
- II = 4th + 5th indices
- 0 = low total of indices

FIGURE 1

## Square 23 of the Chart of the Last Analysis

This is the only square of the hundred into which I divided the chart of the last analysis where there are no misses. Thirty-one of the hundred have more misses than hits.

is to use a finer test, broadening the criteria until negative results are reached.

Rather, that would have been the next step, had Square #23 been typical. A look at the final tabulation (again, see Appendix E) will show that Square #23 is the only one of the hundred that is unanimous in supporting the original theory. The remaining ninety-nine squares range down to a perfectly negative correlation (Square #44, Appendix E), with all the criteria missing the mark. Not even the negative criterion had reliability. By rights, the negative criterion should never have appeared when texts by the same poet were being compared. Yet it did, forty-five out of 295 times, a ratio only slightly better than pure chance. When a text selected more than one other text for positive or negative correlation all pairings were listed. For example text #54 (Browning) correctly chose five Wordsworthian texts in the twenty-third square alone for negative correlation. In each of these five instances all of the indices added up to only ten correspondences. Text #54 selected non-Browning texts sixteen times in 229 trials, but selected another Browning text once. This is only three or four times better than guessing. Likewise the positive criteria, although far better than pure chance, which would score a hit approximately every fourth or fifth time, still did not have the consistency needed for an effective test (see Appendix G).

The present experiment, therefore, is a failure. It did not discover a way of proving authorship by means of words

selected objectively and graphed without regard to text length. Improvements in the method are beyond the scope of the present study, which has now lost some of its innocent objectivity. I have some faith that by working over the data that went into the positive results I can cull out certain of the dead vocabulary items and build a stronger test. It is encouraging that two of the positive criteria, the second and the third, were right roughly four times as often as they were wrong, and that the first criterion was almost sixteen times as effective as guessing; that is, it was right three or four times as often as it was wrong. Especially by concentrating on the relatively few data of Magic Square #23, something viable might be conceived (some of the magic of this square perhaps derives from the fact that seventeen of its twenty-three texts are from Wordsworth, thirteen from The Excursion alone). Future mining of these statistical lodes, however, will have to avoid the pitfall of circular reasoning, and any testing device so derived will itself have to be tested against poems not on my list. Until a test is devised that will work on any hundred out of a hundred texts of known authorship, I, for one, will refuse to accept its results when it is used with texts of doubtful authorship.

## CHAPTER IV

### CONCLUSION: OUTLOOK

The time has come for me to fulfill the promise made in the Introduction to point out the way that the attribution of authorship problem may be solved. To recapitulate: John W. Clark concluded that the problem probably will never be solved, since it is impossible to tell whether close similarities between texts indicates common authorship or mere influence of one author on another. This almost certainly will remain true of disputed texts such as his four or five poems, whose author(s) left no definitely attributable corpora to be used for comparison. Circular arguments, however much pleasure they give the disputant, cannot be said to solve anything; and when one compares data from Gawain with data from Pearl, and data from Patience and Cleanness with data from the other two, all without knowing how many authors are involved or what their known works will yield for data, it seems more than a little vain. If the names of the poems were applied to cigarettes respectively noted for "manly flavour," "taste beyond price," "slower burning," and "less smoke per puff," it still could be that all the brands are rolled in the same shed in Lexington, although not necessarily by the same machine.

Almost as a postscript to the problem taken up in Chapter One comes in 1966 A Concordance to Five Middle English Poems, a computer-derived work by Barnet Kottler and Alan M. Markman

(University of Pittsburgh Press). The concordance relies heavily on the work of Sir Israel Gollancz, whose editions of four of the poems are the basic texts, that have Menner's, Savage's, and Tolkien and Gordon's for variants, and whose edition of the Pearl is one of three variant texts backing up the 1953 Gordon edition. The concordance is a volume that would have saved John W. Clark much of his labor, thereby perhaps depriving us of one of the most truculent pieces of scholarship existing.

In the Kottler and Markman Concordance all the words of the five poems are listed somehow: those in Appendix I are frankly "Partly Concorded," which means that some of the line numbers are given for their occurrences, although why all are not given is not explained. Of special interest here is the decision not to concord a list of 152 words -- a list strange in that it includes words used thousands of times ("the," "and") and words used only once ("foul") -- including all but fifteen of my forty-five test words. The list of 152 merely gives the numbers of occurrences of each item in the five poems as a group. Four more pages added to the book's total of 794 could have presented the numbers of occurrences of these words in each of the five poems. But again the presumption on the part of the compilers is that there will be more interest by authorship scholars in the less common words.

It is nevertheless an improvement over the Matthew Arnold concordance and older concordances generally in which the more common words are merely listed without any tally whatever.

Moreover, at the University of Pittsburgh, according to Kottler and Markman, retrieval of further information is possible, since their input data is kept stored on magnetic tape. But before that tape demagnetizes, it might be best to get everything of possible use in print.

But to return to the recapitulation. Ellegard's conclusion was that the Junius problem had been solved. Against him it can be maintained that circular reasoning was used to show that Francis' style is, not identical with, but closest to that of Junius. Moreover, the data going into his calculations is suspect because it was not objectively or accurately derived. This last objection also is sufficient to vitiate the claims that a project like Milic's might have to validity.

Yet, if there is still anyone who wants to know who the real Junius was, or whether the supposed author of Shakespeare's other works was the same one who wrote The Two Noble Kinsmen, there may be hope. I believe the hope to lie in the method of profiles. If I had to try the project again, I would choose five other writers, and work with only half of their works. I would have a computer count and combine allomorphs for the forty-five items plus a few more: an item for all personal pronouns, and an extra form for "to" would find places in the new list. I would then have all of the operations for deriving my first three indices done by computer.

I would consciously use the circular device to find out which items of vocabulary are most usable in obtaining positive results: wherever "hits" were registered, the profiles would



be scrutinized for points of correspondence. In this way maybe half of the items could be eliminated as having no bearing on the distinctiveness patterns of writers.

Armed with this stripped-down list, I would turn the computer loose on the other half of my writers' works. I would fully expect the assaying power to be great -- much higher than 50%; but I would demand that it test out every time before I would claim for it any efficacy in proving the authorship of doubtful writings.

For if we insist on having recourse to the techniques of science, we must abide by the strict laws of scientific proof. No chemical test is valid unless it always works under controllably identical conditions. If there be an "essence of Keats" underlying subject matter, theme, word order, thought, emotion, and style, that essence ought to be detectible. And if it is truly his essence, then it will be found in every one of his works of appreciable length.

Given the existence of this "essence," "profile," or "handprint," will it ever be possible to feed into a computer a newly-discovered work and get a positive identification or non-identification within seconds of time? I can answer that easy question, "Absolutely yes."

# APPENDIX A

## The Texts Used

Code	Poet	Words in Text	Poem Title	Division
1	T	976	Two Voices	3rd Third
2	T	983	"	2nd Lines
3	T	986	"	3rd Lines
4	T	1004	"	2nd Third
5	T	1007	"	1st Third
6	T	1018	"	1st Lines
7	K	1031	Hyperion	Book III
8	T	1128	Locksley Hall	Odd Lines
9	T	1142	"	Odd Couplets
10	T	1151	"	1st Half
11	W	1175	White Doe of Rylstone	Canto VI
12	W	1269	"	Canto V
13	W	1377	"	Canto IV
14	T	1395	In Memoriam	7th 206 Lines
15	S	1413	Queen Mab	Part II
16	K	1435	The Eve of St. Agnes	2nd Half
17	K	1478	"	1st Half
18	T	1479	Two Voices	2nd Half
19	S	1526	Queen Mab	Part I
20	T	1623	Locksley Hall+60 Years	Even Couplets
21	W	1628	White Doe	Canto I
22	T	1630	Locksley ± 60	1st Half
23	T	1630	"	Odd Lines
24	T	1638	"	Even Lines
25	T	1638	"	2nd Half
26	T	1645	"	Odd Couplets
27	S	1645	Queen Mab	Part VI
28	S	1652	"	Part III
29	S	1682	"	Part VIII
30	S	1732	"	Part IX
*31	K	1747	Isabella	
32	S	1761	Daemon of the World	Part I
33	S	1834	Queen Mab	Part V
34	W	1896	Michael	1st Half
35	S	1922	Queen Mab	Part VII
36	S	1968	"	Part IV
37	T	1989	The Princess	Part I
38	T	1991	Oenone	
39	W	1992	Michael	2nd Half
40	K	2035	The Fall of Hyperion	2nd Half
41	K	2037	"	1st Half
42	W	2055	The Borderers	Act IV
43	S	2062	Lines Written Among the Euganean Hills	
44	B	2062	Pippa Passes	Night
45	B	2075	"	Noon

Code	Poet	Words in Text	Poem Title	Division
46	W	2086	White Doe	Canto I
47	W	2114	The Excursion	1st Half, Book VIII
48	W	2160	"	1st Third, Book II
49	W	2173	White Doe	Canto VII
50	W	2193	"	Canto III
51	S	2199	Mask of Anarchy	
52	W	2219	The Excursion	2nd Half, Book VIII
53	T	2254	Locksley Hall	
54	B	2268	Pippa Passes	Morning
55	W	2270	The Excursion	3rd Third, Book II
56	S	2278	Hellas	Rimed Lines
57	W	2279	The Prelude	2nd Half, Book III
58	W	2295	The Excursion	2nd Third, Book II
59	W	2316	"	3rd Fourth, Book IV
60	W	2332	The Prelude	1st Half, Book III
61	W	2334	The Excursion	3rd Third, Book III
62	W	2339	"	1st Fourth, Book IV
63	K	2357	Otho the Great	Act II
64	W	2360	The Excursion	2nd Third, Book III
65	W	2363	"	2nd Fourth, Book IV
66	W	2366	"	1st Third, Book I
67	W	2368	"	4th Fourth, Book IV
68	W	2375	"	1st Third, Book V
69	K	2375	Lania	Part II
70	S	2386	The Cenci	1st Half, Act V
71	S	2387	Daemon of the World	Part II
72	S	2397	The Sensitive Plant	
73	W	2405	The Excursion	1st Third, Book III
74	S	2449	The Cenci	2nd Half, Act V
75	W	2452	The Excursion	2nd Third, Book I
76	W	2456	The Prelude	Book XII
77	W	2463	The Excursion	2nd Third, Book V
78	T	2471	Pelleas and Ettarre	1st Half
79	T	2477	Balin and Balan	1st Half
80	W	2485	The Excursion	3rd Third, Book V
81	S	2491	Letter to Maria Gisborne	
82	S	2506	Hellas	1st Third
83	K	2530	Otho the Great	Act IV
84	T	2535	Pelleas and Ettarre	2nd Half
85	T	2535	Balin and Balan	2nd Half
86	W	2575	The Excursion	1st Third, Book VII
87	W	2584	"	2nd Third, Book VII
88	W	2585	"	3rd Third, Book VII
89	W	2585	"	3rd Third, Book I
90	T	2615	Merlin and Vivian	1st Third
91	T	2627	"	3rd Third
92	S	2634	Hellas	3rd Third
93	K	2635	Otho the Great	Act III

Code	Poet	Words in Text	Poem Title	Division
94	S	2633	Hellas	2nd Third
*95				
96	T	2674	In Memoriam	6th 412 Lines
97	T	2685	Merlin and Vivian	2nd Third
98	T	2686	In Memoriam	2nd 412 Lines
99	W	2693	Guilt and Sorrow	2nd Half
100	T	2702	In Memoriam	5th 412 Lines
101	B	2704	The Pope (Ring and Book)	4th Sixth
102	B	2707	"	2nd Sixth
103	T	2709	In Memoriam	4th 412 Lines
104	K	2724	Hyperion	Book I
105	S	2725	Charles the First	1st Half
106	S	2725	"	2nd Half
107	W	2729	Musings Near Aquapendente	
108	W	2731	Guilt and Sorrow	3 Rimes
109	B	2734	The Pope	1st Sixth
110	T	2736	In Memoriam	3rd 412 Lines
111	B	2738	The Pope	5th Sixth
112	B	2749	"	6th Sixth
113	T	2752	In Memoriam	1st 412 Lines
114	W	2754	Guilt and Sorrow	1st Half
115	T	2760	In Memoriam	7th 412 Lines
116	W	2767	The Prelude	Book XIII
117	B	2775	Guisseppe Caponsacchi (R & B)	5th Sixth
118	T	2778	The Princess	Part III
119	B	2805	The Pope	3rd Sixth
120	B	2810	Guisseppe Caponsacchi	6th Sixth
121	B	2812	"	2nd Sixth
122	B	2817	"	1st Sixth
*124	S	2823	The Cenci	Act II
125	T	2839	The Princess	Part VII
126	W	2860	The Excursion	1st Third, Book VI
127	K	2860	Otho the Great	Act V
128	W	2871	The Excursion	2nd Third, Book VI
129	W	2878	"	1st Half, Book IX
130	W	2886	"	2nd Half, Book IX
131	B	2886	Guisseppe Caponsacchi	4th Sixth
132	B	2898	"	3rd Sixth
133	T	2905	The Lover's Tale	3rd Fourth
134	T	2910	"	1st Fourth
135	T	2915	"	2nd Fourth
136	B	2922	Half Rome (Ring and Book)	2nd Fourth
137	K	2975	Hyperion	Book II
138	B	2976	Half Rome	1st Fourth
139	T	2987	Two Voices	
140	B	2998	Half Rome	3rd Fourth
141	W	3004	The Idiot Boy	
142	B	3011	Half Rome	4th Fourth
143	T	3037	The Lover's Tale	4th Fourth

Code	Poet	Words in Text	Poem Title	Division
144	K	3056	Lamia	Part I
145	T	3060	The Princess	Part VI
146	S	3061	Prometheus Unbound	4th 400 Blanks
147	T	3067	The Last Tournament	1st Half
148	S	3071	Prometheus Unbound	3rd 400 Blanks
149	K	3077	Sleep and Poetry	
150	T	3114	The Last Tournament	2nd Half
151	S	3131	Prometheus Unbound	1st 400 Blanks
152	S	3150	"	2nd 400 Blanks
153	T	3268	Locksley Hall Sixty Years After	
154	W	3321	The Prelude	Book XIV
155	S	3323	The Cenci	Act I
156	S	3334	Queen Mab	Parts III & VIII
157	S	3335	"	Parts II & VII
158	T	3415	The Marriage of Geraint	1st Half
159	T	3444	"	2nd Half
160	W	3444	An Evening Walk	
161	W	3464	The Prelude	Book XI
162	W	3470	"	Book II
163	T	3501	The Ring	
164	W	3523	The Prelude	Book IV
165	W	3553	The Brothers	
166	T	3567	The Princess	Part II
167	K	3603	Endymion	1st Half, Book I
168	K	3622	"	1st Half, Book IV
169	B	3687	Pompilia (Ring and Book)	2nd Fourth
170	B	3739	"	3rd Fourth
171	K	3751	Endymion	2nd Half, Book I
172	T	3755	Gareth and Lynette	2nd Third
173	S	3761	Prometheus Unbound	Act III
**				
**				
**				
**				
177	K	3777	Otho the Great	Act I
178	T	3779	Gareth and Lynette	3rd Third
179	K	3798	Endymion	1st Half, Book II
180	B	3801	Pompilia	4th Fourth
181	S	3811	The Cenci	Act III
182	T	3812	Gareth and Lynette	1st Third
183	B	3820	Pompilia	1st Fourth
184	K	3826	Endymion	2nd Half, Book III
185	S	3857	Prometheus Unbound	Act I
186	T	3868	The Passing of Arthur	
187	K	3872	Endymion	2nd Half, Book II
188	S	3888	The Cenci	Act IV
189	K	3892	Endymion	2nd Half, Book IV
190	T	3925	Geraint and Enid	1st Half
191	K	3961	Endymion	1st Half, Book III
192	S	3968	Adonais	
193	T	4025	Geraint and Enid	2nd Half

Code	Poet	Words in Text	Poem Title	Division
194	W	4243	The Borderers	Act III
195	W	4266	The Prelude	Book IX
196	T	4275	The Coming of Arthur	
197	S	4304	The Triumph of Life	
*				
199	T	4407	The Princess	Part V
200	W	4433	The Prelude	Book X
201	S	4492	Rosalind and Helen	1st Half
202	S	4547	"	2nd Half
203	T	4562	The Princess	Part IV
*				
205	W	4574	The Prelude	Book V
206	W	4583	The Borderers	Act I
207	T	4672	In Memoriam	2nd Half, Even Lines
208	T	4690	"	2nd Half, 3 Rimes
209	T	4744	"	1st Half, Even Lines
210	T	4760	"	2nd Half, A Rimes
211	T	4761	"	1st Half, 3 Rimes
212	W	4770	The Prelude	Book I
213	T	4778	In Memoriam	2nd Half, Odd Lines
*				
215	T	4828	In Memoriam	1st Half, A Rimes
216	T	4845	"	1st Half, Odd Lines
217	W	4875	The Borderers	Act II
218	S	4922	Julian and Maddalo	
219	W	4983	The Prelude	Book VIII
220	S	5028	Peter Bell the Third	
221	W	5107	Descriptive Sketches	
222	S	5359	Swellfoot the Tyrant	
223	S	5392	Alastor	
224	T	5411	In Memoriam	2nd 824 Lines
225	T	5512	"	1st & 7th 412 Lines
226	W	5525	The Prelude	Book VII
227	W	5659	"	Book VI
228	T	5727	Guinevere	
229	T	5889	Lancelot and Elaine	1st Half
230	T	5919	"	2nd Half
231	K	6138	The Cap and Bells	
232	W	6151	Tour of the Alps	
233	T	6613	Aylmer's Field	
234	W	7175	Peter Bell	
235	T	7229	Enoch Arden	
236	T	7586	The Holy Grail	
237	B	7951	Bishop Blougram's Apology	

\* Text #31, Isabella, suffers from defective computer print-out. It was retained anyway, as a control.

\* Eight texts were temporarily misplaced when the charts were being made.

~~\*\*\*~~ In the listing on the preceding pages, the column headed Division has arabic numbers when the division is mine; roman numerals indicate the divisions I found in the poems. The original project planned was to make much of comparing within poems to see whether, for example, similarities were greater between the first and second halves of a poem than between its odd and even lines. Such divisions are magnificently easy to accomplish with IBM cards, as are the subsequent combinations of glossaries. The computer, just like anybody else, is able to alphabetize fifty words much more than twice as fast as one hundred. Dividing the poems, therefore, not only gave greater flexibility to the project, but it also saved computer time.

The poems on this list were, for the most part, key-punched from the following editions:

Poems of Robert Browning, ed. Donald Smalley, Houghton Mifflin, 1956.

Complete Poems of Keats and Shelley, Modern Library, Random House, n. d.

Poems of Tennyson, ed. Jerome H. Buckley, Houghton Mifflin, 1956.

The Poetical Works of Alfred, Lord Tennyson, Hins and Knight, Troy, N. Y., 1937.

Poetical Works of Wordsworth, ed. Thomas Hutchinson, Oxford University Press, 1904 (1960).

## APPENDIX B

### The Thirteen Texts

From Which were Derived the Forty-five Test Words

<u>Text</u>	<u>Division</u>	<u>Number of Words</u>
Endymion	2nd Half of Part II, 1st Half of Part III	7833
Geraint and Enid		7950
Bishop Bloughram's Apology		7951
The Cenci	Acts I and V	8158
The Prelude	Books X and XI	7897
Hellas		7778
Merlin and Vivian		7927
The Excursion	2nd Half of Book VIII, Book IX	7983
Otho the Great	Acts III, IV, V	8025
The Pope (The Ring and the Book)	2nd Half	8191
The Princess	Parts I, II, III	8334
In Memoriam	1st 1236 Lines	8174
In Memoriam	2nd 1236 Lines	8085

Note the discrepancy in word population between the last two items. It often happens that the same number of verses will contain widely different numbers of words. A statistician would consider the word to be the unit, but the poet would more likely think of the single line of verse as the unit. Lines of regular verse, moreover, are more "like" one another than are individual words, which range from the zero-like content of an initial "it" or "there" to the most connotative abstract or concrete terms.

The project I finally determined upon ignores both measures of size by touching only non-content words and discarding all others. Herein is some of my justification for not distinguishing text size during the main portion of the project.



## APPENDIX C

### The Forty-five Test Words with the Allomorphs of Fourteen

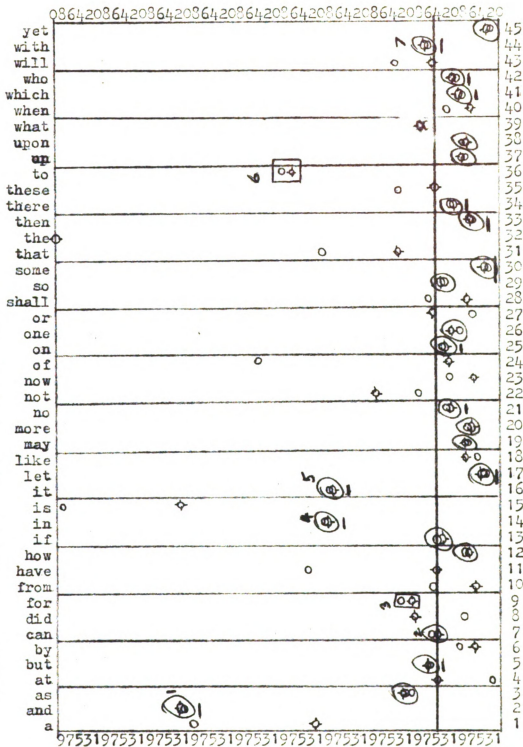
<u>Word</u>	<u>Allomorphs</u>
a	an
and	
as	
at	
but	
by	
can	cannot, cannt (can't), could
did	do, done, dost, doth
for	
from	
have	had, hadst, hath, has, hast, having
how	
if	
in	into
is	am, are, art, be, been, being, isit, 'tis, 'twas, 'twere, was, wast, were, wert
it	isit, 'tis, 'twas, 'twere
let	
like	likes
may	mayst, might, mightst
more	
no	
not	cannot, cannt
now	
of	
on	
one	
or	
shall	shalt, should, shouldst
so	
some	
that	
the	
then	
there	
these	this
to	
up	
upon	
what	
when	
which	
who	whom, whose
will	wilt, would, wouldst
with	
yet	

### Comparison of the Graphs for Texts 194 (o) and 237 (♦)

The circled  
pairs correspond  
(1). Those un-  
defined corres-  
pond closely (2).  
There are seven  
correspondences  
above the 15%  
line (3). When  
the 237 graph  
is moved up  
there are two  
additional cor-  
respondences  
above the 15%  
line (4). Seven  
of these are  
close (5).

Thus the five indices for the comparison of texts 194 & 237 are:

(1) (2) (3) (4) (5)  
24, 13, 7, 9, 7.



## APPENDIX E

### A Summation of the Chart of the Last Analysis

1	17/5	3/2	12/3	10/7	9/7	12/7	13/7	8/4	10/2	5/1
2	19/8	16/4	14/5	12/5	10/4	8/2	11/1	12/2	11/4	13/2
3	16/7	11/6	21/0	16/2	8/3	26/7	10/5	7/3	9/2	8/2
4	14/6	17/4	16/4	14/6	16/3	13/5	14/3	11/2	21/3	22/5
5	13/7	8/5	10/5	3/3	17/2	11/7	4/3	7/4	4/2	9/3
6	13/8	13/9	13/3	13/8	15/7	17/3	7/5	9/3	5/1	6/5
7	12/8	21/4	11/2	14/4	14/8	10/3	11/3	17/8	6/2	12/6
8	13/12	12/4	11/3	12/5	14/7	16/5	20/7	19/9	7/3	9/3
9	11/5	14/5	17/9	20/6	15/4	9/4	13/6	14/5	16/4	26/3
10	11/9	13/9	10/6	20/6	17/6	14/10	19/8	13/4	20/6	14/4

Each of the "improper fractions" above represents the score for a square consisting of 23 x 23 smaller squares. Within most squares there is a possibility of as many as 1,587 indices of correspondence or of 529 indices of non-correspondence. The full chart is 230 by 230 squares or twelve by twelve feet.

The top figure in each square above is the number of criteria pertaining to the more than five hundred pairings of graphs represented by each square, and the lower number indicates the times the criteria were mistaken.

See Table 1, page 42, for a close view of one portion of this chart. Table 1 represents only 1% of the total chart; it is anything but typical of my results.

## APPENDIX F

### Specimens from Different Stages of the Project

(1) Program cards, the first seven of 115 in the deck of the word count program built by James D. Clark of the Psychology Department, Michigan State University.

(2) Input cards for this project. Each card contains one line of verse, every word verified and, in some instances, respelled according to a convention (see pages 34-5). The punching of the cards was done almost wholly in my home, on a machine graciously rented to me by IBM. Blank cards were purchased from the Computer Center at Michigan State University.

(3) Output sheet, to show how the computer completes its part of the project.

(4) Listing sheet with the 45 test words in various possible forms.

(5) Gathering sheet, with the 45 items totalled.

(6) Broken-line graph, in which the broken line does not really mean anything, but merely takes the eye from point to point (however, the broken line is the profile of a text).

(7) Same graph with the points expanded to  $\frac{1}{4}$ " holes. This graph, and 229 others, were the core of my project; comparing each one with all the others consumed the better part of a full year. The same job done by computer (that is, what turned out to be the effective part of it) would have taken less than two weeks, including all preparation.

(8) One leaf from the 12' x 20' chart containing the indices derived from comparing graphs (see pages 38-9).

**Specimen 1**

READ 3,ND

KU = 5000

[illegible]

CON(KT=2000000790000000R)

COMMON NAME

[illegible]

```

DIMENSION LEAD(30),FY(10),NAM(12000),NUM(4),MUN(1F,4),

```

PROGRAM MC

iii i i

000000000000000000200090806C0000000000000C0000000000000000000000000000000000

541085 ZIMMER

iii

IC, OP, S4

MSU COMPUTER LABORATORY IDENTIFICATION CARD

LL

[.]

# M

[illegible]

**S**

2 U

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80  
BSC 655859

**NSC 653660**

IN THE SAME CIRCLE WE BELIEVE  
 THE DOCKET WOULD REST I DARE NOT DOUBT

314

THE DOCKET WOULD REST I DARE NOT DOUBT  
 OR THIS WOULD BE THE SAME

313

OR THIS WOULD BE THE SAME

WITH THREE WOULD BE THE SAME  
 THE DOCKET WOULD REST I DARE NOT DOUBT

312

THE DOCKET WOULD REST I DARE NOT DOUBT  
 THE DOCKET WOULD REST I DARE NOT DOUBT

310

THE DOCKET WOULD REST I DARE NOT DOUBT  
 THE DOCKET WOULD REST I DARE NOT DOUBT

301

2 voices

last 3rd of 1950

FOR COMMENT		FORTRAN STATEMENT		IDENTIFICATION	
STATEMENT NUMBER	CHARACTER	STATEMENT	CHARACTER	STATEMENT	CHARACTER
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0
6	0	0	0	0	0
7	0	0	0	0	0
8	0	0	0	0	0
9	0	0	0	0	0
10	0	0	0	0	0
11	0	0	0	0	0
12	0	0	0	0	0
13	0	0	0	0	0
14	0	0	0	0	0
15	0	0	0	0	0
16	0	0	0	0	0
17	0	0	0	0	0
18	0	0	0	0	0
19	0	0	0	0	0
20	0	0	0	0	0
21	0	0	0	0	0
22	0	0	0	0	0
23	0	0	0	0	0
24	0	0	0	0	0
25	0	0	0	0	0
26	0	0	0	0	0
27	0	0	0	0	0
28	0	0	0	0	0
29	0	0	0	0	0
30	0	0	0	0	0
31	0	0	0	0	0
32	0	0	0	0	0
33	0	0	0	0	0
34	0	0	0	0	0
35	0	0	0	0	0
36	0	0	0	0	0
37	0	0	0	0	0
38	0	0	0	0	0
39	0	0	0	0	0
40	0	0	0	0	0
41	0	0	0	0	0
42	0	0	0	0	0
43	0	0	0	0	0
44	0	0	0	0	0
45	0	0	0	0	0
46	0	0	0	0	0
47	0	0	0	0	0
48	0	0	0	0	0
49	0	0	0	0	0
50	0	0	0	0	0
51	0	0	0	0	0
52	0	0	0	0	0
53	0	0	0	0	0
54	0	0	0	0	0
55	0	0	0	0	0
56	0	0	0	0	0
57	0	0	0	0	0
58	0	0	0	0	0
59	0	0	0	0	0
60	0	0	0	0	0
61	0	0	0	0	0
62	0	0	0	0	0
63	0	0	0	0	0
64	0	0	0	0	0
65	0	0	0	0	0
66	0	0	0	0	0
67	0	0	0	0	0
68	0	0	0	0	0
69	0	0	0	0	0
70	0	0	0	0	0
71	0	0	0	0	0
72	0	0	0	0	0
73	0	0	0	0	0
74	0	0	0	0	0
75	0	0	0	0	0
76	0	0	0	0	0
77	0	0	0	0	0
78	0	0	0	0	0
79	0	0	0	0	0
80	0	0	0	0	0
81	0	0	0	0	0
82	0	0	0	0	0
83	0	0	0	0	0
84	0	0	0	0	0
85	0	0	0	0	0
86	0	0	0	0	0
87	0	0	0	0	0
88	0	0	0	0	0
89	0	0	0	0	0
90	0	0	0	0	0
91	0	0	0	0	0
92	0	0	0	0	0
93	0	0	0	0	0
94	0	0	0	0	0
95	0	0	0	0	0
96	0	0	0	0	0
97	0	0	0	0	0
98	0	0	0	0	0
99	0	0	0	0	0
100	0	0	0	0	0

VABCO 680127 SP

PROPLE 1 8  
TWO VOICES 3RD THIRD

1. A	20	2. ABOVE	1	3. AEOLIAN	1	4. AGAIN	3
5. AGAINST	1	6. AIR	1	7. AIRS	1	8. ALL	3
9. ALONE	1	10. ALONG	1	11. ALPINE	1	12. ALSO	1
13. ALTHOUGH	1	14. ALWAYS	1	15. AN	2	16. ANCHOR	1
485. WORDS	1	486. WORDS	2	487. WORTH	1	488. WOULD	2
489. WRECK	1	490. WRINKLES	1	491. WRONG	1	492. WROUGHT	1
493. YEAR	1	494. YEARNING	1	495. YEARS	1	496. YET	1
497. YOU	1						

Specimen 3

TOTAL NUMBER OF WORDS = 976





20	if 6	these 2
me 0	in 16	this 3
me 2	into 1	tin 1
ml 25	is 8	to 17
me 2	it 0	twice 0
it 1	it 5	twice 0
rs 8	its 1	twould 0
st 4	let 0	up 1
be 4	like 7+1	upon 1
been 0	may 3	was 5
beings 0	might 6+1	was 0
but 12	more 2	were 2
by 2	no 7	work 2
can 1+1	not 9	what 2
can 1	now 1	when 3
can 2	of 22	which 5
can 0	on 4	who 3
can 1	one 5	where 0
can 0	or 3	where 3
one 1	shall 0	will 0
with 0	shall 0	will 1
or 8	should 2	with 16
was 8	so 6	would 2
ad 0	some 5	would 0
as 1+1	that 20	yet 1
both 0	the 40	0
me 0	then 4	
living 0	there 3	
how 2		

Specimen 4

## Specimen 5

a	22	20+2	a	20+2
and	25		and	25
as	8		as	8
at	4		at	4
but	12		but	12
by	2		by	2
can	5	2+1+2	can	2+1+2
did	2	1+1	did	1+1
for	8		for	8
from	8		from	8
have	2		have	2
how	2		how	2
if	6		if	6
in	17	16+1	in	16+1
is	25	2+1+4+8+1+5+2+2	is	2+1+4+8+1+5+2+2
it	7	5+1+1	it	5+1+1
let	0		let	0
like	8		like	8
may	10	3+7	may	3+7
more	2		more	2
no	7		no	7
not	10	1+9	not	1+9
now	1		now	1
of	22		of	22
on	4		on	4
one	5		one	5
or	3		or	3
shall	2		shall	2
so	6		so	6
some	5		some	5
that	20		that	20
the	40		the	40
then	4		then	4
there	3		there	3
these	5	3+3	these	2+3
to	17		to	17
up	1		up	1
upon	1		upon	1
what	2		what	2
when	3		when	3
which	5		which	5
who	6	3+3	who	3+3
will	3	1+2	will	1+2
with	16		with	16
yet	1		yet	1

1-T-976

40

20

72

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65  
66  
67  
68  
69  
70  
71  
72  
73  
74  
75  
76  
77  
78  
79  
80  
81  
82  
83  
84  
85  
86  
87  
88  
89  
90  
91  
92  
93  
94  
95  
96  
97  
98  
99  
100

1  
T  
976

Specimen 6

26

24

22

20

18

16

14

12

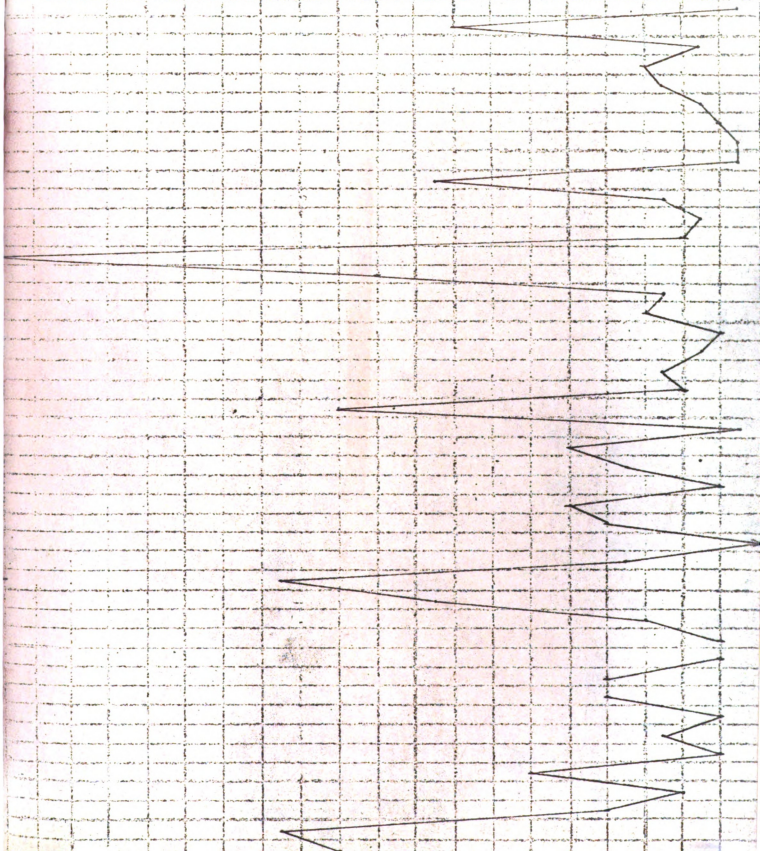
10

8

6

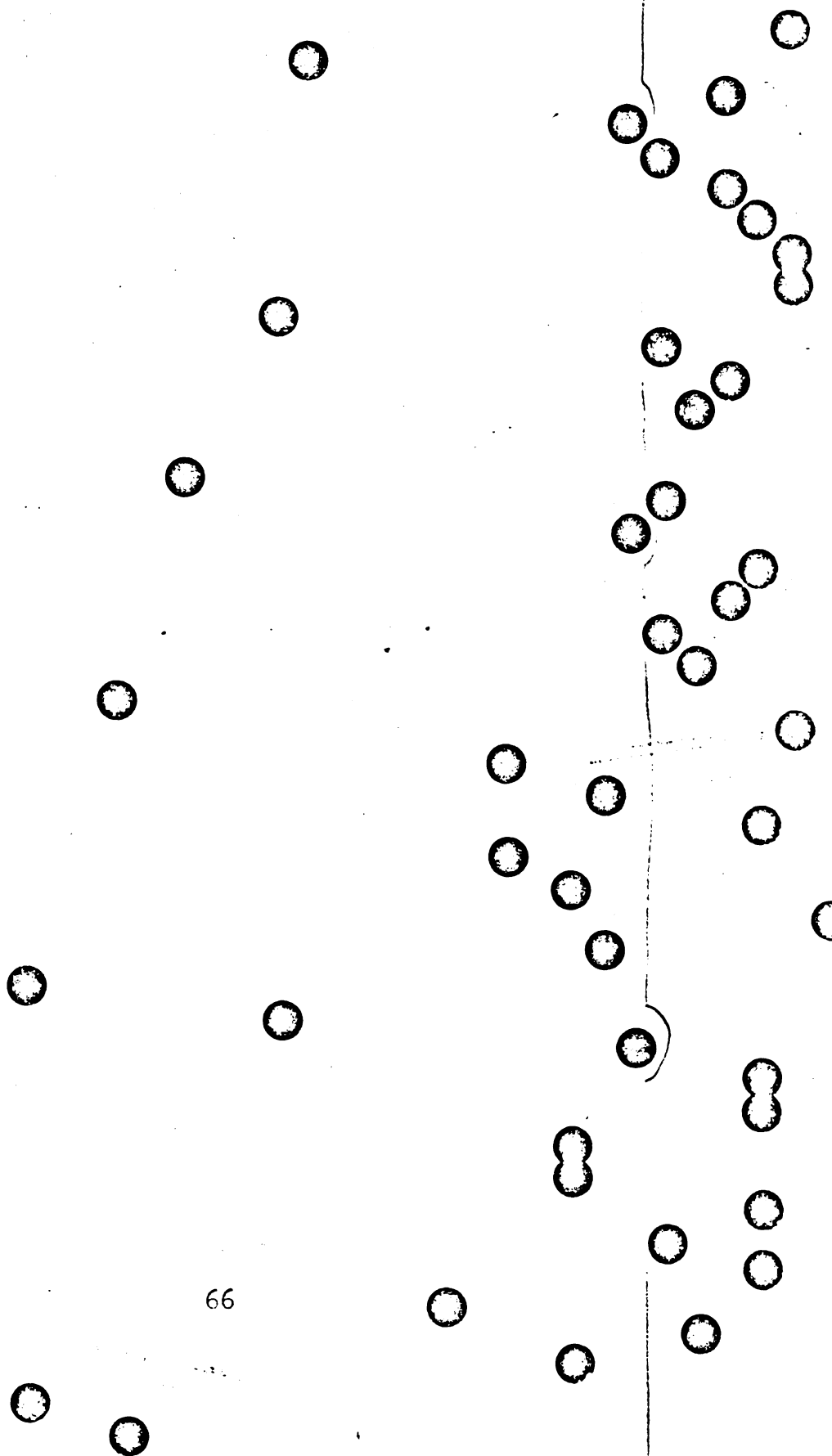
4

2



-1

Specimen 7



#51

II-13

#52

1, 8, 9, <sup>4.3</sup><sub>5.2</sub>

♂-10

#53

1, 17, 1, 3, 2 26, 13, 3, <sup>3.3</sup><sub>3.3</sub>

♂-8

#54

2, 2, 2, 2

4, 2, 0, 3, 1

6, 3, 1, 3, 1

♂-20

#55

3, 11, 3, 4, 3

20, 6, 1, <sup>3.2</sup><sub>4.3</sub>13, 6, 0, <sup>1.0</sup><sub>3.1</sub>9, 5, 4, <sup>4.4</sup><sub>4.4</sub>

♂-16

3, 11, 1, <sup>1.1</sup><sub>3.1</sub>25, 15, 5, <sup>5.3</sup><sub>5.3</sub>23, 12, 2, <sup>2.2</sup><sub>2.2</sub>

8, 3, 1, 1, 1

10, 4, 1, 6.4

1, 5, 4, 9, 3

17, 8, 1, <sup>3.3</sup><sub>7.5</sub>10, 4, 1, <sup>1.0</sup><sub>2.1</sub>16, 5, 5, <sup>6.5</sup><sub>6.5</sub>16, 9, 7, <sup>7.4</sup><sub>7.5</sub>

2, 4, 2, 5, 4

17, 5, 2, <sup>4.4</sup><sub>6.5</sub>9, 1, 0, <sup>0.0</sup><sub>1.1</sub>

12, 4, 2, 5.4

20, 9, 6, <sup>6.5</sup><sub>8.1</sub>

10, 10, 1, 5, 4

32, 16, 5, <sup>6.3</sup><sub>6.4</sub>19, 12, 3, <sup>3.2</sup><sub>4.2</sub>10, 1, 2, <sup>2.2</sup><sub>4.1</sub>15, 7, 3, <sup>4.3</sup><sub>4.3</sub>2, 6, 3, <sup>3.3</sup><sub>3.3</sub>15, 6, 1, <sup>2.0</sup><sub>5.1</sub>8, 5, 0, <sup>0.0</sup><sub>1.1</sub>13, 2, 4, <sup>4.2</sup><sub>4.2</sub>16, 8, 5, <sup>5.4</sup><sub>5.4</sub>

12, 5, 1, 4, 2

23, 10, 1, <sup>3.2</sup><sub>7.3</sub>16, 6, 2, <sup>2.1</sup><sub>3.1</sub>

11, 6, 0, 4.2

21, 12, 5, <sup>8.5</sup><sub>8.5</sub>15, 6, 4, <sup>4.3</sup><sub>4.3</sub>19, 9, 1, <sup>3.2</sup><sub>7.4</sub>12, 6, 0, <sup>1.1</sup><sub>2.1</sub>

10, 4, 1, 7.4

15, 5, 4, <sup>4.4</sup><sub>4.4</sub>

## APPENDIX G

### The Effectiveness of the Four Criteria

Criterion*	<u>Hits</u> Misses	Ratio	Expectation	Effectiveness
2	222/67**	3.3	.2	16.5
3	238/233	1.02	.2	5.1
II	184/168	1.1	.2	5.5
0	247/47	5.25	5.0	1.05

\*For an explanation of the symbols 2, 3, II, and 0, see p. 40.

\*\*The figures in this column are reasonably accurate. When I made the count last year the fractions were: 222/68, 236/233, 189/162, and 250/45.

Since the texts behind the graphs used in this project came from five different poets, every fifth time two graphs were compared the texts behind them were by the same poet. A person guessing blindly the identity of the second poet in each case could expect to be correct one out of every five times (more often with Wordsworth, less often with Keats). Thus the expected ratio is one out of five ( $1/5$ ), or .2.

Yet the positive criteria (2, 3, and II) guessed right five to sixteen times as often as a blind guesser would. Therefore the positive criteria can be said to be five to sixteen times as effective as sheer guessing.

The negative criterion (0) is obviously worthless as an indicator of identity since its effectiveness quotient is practically the same as the quotient for chance guessing.

The positive criterion (2) shows some promise. With certain testing items removed -- I have no notion which ones -- and others added, and with a confining of the testing to texts of more nearly equal size, the effectiveness quotient could no doubt be raised considerably. Sixteen may be high enough for some purposes, as when only two possibilities are present in an authorship dispute. One hundred would be better when more potential authors are in the competition. But when almost anyone could have written a doubtful text even a quotient of infinity would not be a positive identification.

## SOURCES CITED

Clark, John Williams. The Authorship of Sir Gawain and the Green Knight, Pearl, Cleanness, Patience, and Erkenwald in the Light of the Vocabulary. University of Minnesota doctoral dissertation, unpublished, 1941.

### Cited from Clark:

Bateson, Hartley, ed. Patience. Second Edition. London, 1913.

Chambers, R. W. "Long Will, Dante, and the Righteous Heathen," Essays and Studies by Members of the English Association, IX (1924), 50.

Chapman, Coolidge Otis. "The Authorship of the Pearl," Publications of the Modern Language Association, XXXVII (1932), 346.

-----, A Lexical Concordance of the Middle English Pearl, Cleanness, Patience, and Sir Gawayne and the Grene Knight. A--, K--, No-- - Wy--, and Z-- only. Cornell University doctoral thesis. 1927.

Gollancz, Sir Israel, ed. Cleanness. Volume I, Introduction, Text, and Notes. London, 1921.

-----, and Mabel Day, edd. Cleanness. Volume II, Glossary. London, 1933.

-----, ed. Patience. London, 1913.

-----, ed. Pearl. Revised edition. London, 1921.

-----, ed. St Erkenwald. London, 1922.

Horstmann, Carl, ed. Altenglische Legenden (Neue Folge). Heilbrunn, 1881.

Oakden, J. P. Alliterative Poetry in Middle English. Volume I, The Dialectal and Metrical Survey. Publications of the University of Manchester, CCV (1930).

-----, -----, Volume II, A Survey of the Tradition. Publications of the University of Manchester, CCXLVI (1935).

Savage, Henry L., ed. St Erkenwald. New Haven, 1926.

Serjeantson, Mary S. "The Dialects of the West Midlands in Middle English," Review of English Studies, III (1927), 54, 186, 319.

Ten Brink, Bernhard. History of English Literature.  
Volume I. Eng. tr., H. L. Kennedy. New York, 1903.

Tolkien, J. R. R. and E. V. Gordon, edd. Sir Gawain and the Green Knight. London, 1925.

Trautmann, Moritz. Ueber Verfasser und Entstehungszeit einiger Alliterierender Gedichte des Altenglischen.  
Halle, 1876.

Conley, John. Review of H. L. Savage's The Gawain-Poet: Studies in his Personality and Background. Chapel Hill, 1956. In Speculum, XXII (1957), 858-61.

Ellegard, Alvar. Who Was Junius? Stockholm, 1962.

-----, A Statistical Method for Determining Authorship.  
Stockholm, 1962.

Reviews of Ellegard:

"The Statistics of Style," Times Literary Supplement. Jan. 5, 1963, page 1.

Zimmer, George W. Journal of English and Germanic Philology, LXII (1963), 618-9.

Herdan, Gustave. Type-Token Mathematics: A Textbook of Mathematical Linguistics. 's-Gravehage, 1960.

Kottler, Barnot, and Alan M. Markman. A Concordance to Five Middle English Poems. Pittsburgh, 1966.

Miles, Josephine. "Eras in English Poetry," Publications of the Modern Language Association, LXX (1955), 655-75.

Milic, Louis T. A Quantitative Approach to the Style of Jonathan Swift. The Hague, 1967.

-----, Dissertation Abstracts, XXIV (1964), 3730.

Mosteller, Frederick, and David L Wallace. Inference and Disputed Authorship: The Federalist. Reading, Mass., 1964.

Taylor, George C. "Montaigne-Shakespeare and the Deadly Parallel," Philological Quarterly, XXII (1943), 330-3.

Yule, George Udny. The Statistical Study of Literary Vocabulary. Cambridge, 1944.

-----, "On Sentence-length as a Statistical Characteristic of Style," Biometrika, XXV (1938), 363-90.



MICHIGAN STATE UNIV. LIBRARIES



31293010631491