R E P O R T   R E S U M E S

ED 018 407 TE 000 290

STEPHEN CRANE'S "THE O'RUDDY"--A PROBLEM IN AUTHORSHIP DISCRIMINATION.

BY- O'DONNELL, BERNARD

PUB DATE 66

EDRS PRICE MF-$0.25 HC-$0.48 10P.

DESCRIPTORS- *COMPOSITION (LITERARY), *DISCRIMINANT ANALYSIS, *ENGLISH INSTRUCTION, *PROSE, LANGUAGE, LANGUAGE PATTERNS, COMPARATIVE ANALYSIS, ELECTRONIC DATA PROCESSING, RHETORIC, SEMANTICS, DICTION, SYNTAX, GRAMMAR, SENTENCE STRUCTURE, LITERATURE, STEPHEN CRANE, ROBERT BARR, "THE O'RUDDY",

THE PURPOSE OF THIS ANALYSIS WAS TO DISCOVER CERTAIN ASPECTS OF STYLE (BOTH LEXICAL AND GRAMMATICAL) WHICH COULD BE COUNTED AND WHICH WOULD, WHEN COMPARED, DIFFERENTIATE BETWEEN THE WRITTEN PROSE OF TWO AUTHORS. THE SUBJECT SELECTED FOR ANALYSIS WAS "THE O'RUDDY," BEGUN BY STEPHEN CRANE AND COMPLETED BY ROBERT BARR. SINCE THERE WAS NO RECORD OF THE POINT OF TRANSITION, THE AUTHORSHIP OF VARIOUS CHAPTERS WAS IN DOUBT. THE RESULTS OF A PILOT STUDY, DESIGNED TO GENERATE TECHNIQUES FOR SUCH WORK, INDICATED THAT NOT ONLY SINGLE ITEMS, BUT ALSO PATTERNS OF INTERRELATIONSHIPS WOULD DISCRIMINATE. AFTER PRELIMINARY ANALYSIS TO GENERATE 18 WEIGHTED VARIABLES, AND A SUBSEQUENT CHECK ON THE VALIDITY OF THOSE VARIABLES AND THEIR WEIGHTING, THE NEW TECHNIQUE WAS APPLIED TO PARAGRAPHS, BLOCKS OF 1,000 WORDS, AND WHOLE CHAPTERS OF "THE O'RUDDY," RESULTING IN THE ASSIGNMENT OF CHAPTERS 1-24 TO CRANE, CHAPTER 25 AS TRANSITIONAL, AND CHAPTERS 26-33 TO BARR. THIS RESEARCH METHOD OFFERS A WAY OF ANALYZING NOT ONLY THE PATTERN OF STUDENT COMPOSITION, BUT ALSO THE PATTERN AND SPECIFIC STYLISTIC DEVICES OF ACCEPTED AUTHORS. THE RESULTS OF THIS STUDY HIGHLIGHT THE IMPORTANCE OF THE VARIABLES IN THE COMPOSING PROCESS. (THIS ESSAY APPEARED IN "THE COMPUTER AND LITERARY STYLE," KENT STUDIES IN ENGLISH, NUMBER 2, KENT STATE UNIVERSITY PRESS, OHIO, 1966. IT IS BASED ON A DOCTORAL DISSERTATION COMPLETED AT HARVARD UNIVERSITY, 1963.) (MM)

Stephen Crane's *The O'Ruddy*: A Problem in Authorship Discrimination

BERNARD O'DONNELL, *University of Iowa*

REPRINTED FROM

THE COMPUTER & LITERARY STYLE

KENT STUDIES IN ENGLISH, NUMBER 2

STEPHEN CRANE'S THE O'RUDDY: A PROBLEM IN AUTHORSHIP DISCRIMINATION

BERNARD O'DONNELL

CASES OF doubtful or disputed authorship are not new to literary history. The Shakespeare controversy, for example, has been with us, with intermittant boiling points, for over two hundred years. The authorship of *The Imitation of Christ,* now generally attributed to Thomas à Kempis, was disputed for several hundred years. In the past, literary historians advanced tentative hypotheses based on stylistic and historical data; these hypotheses were often disputed by other literary historians. Conclusions were only rarely agreed upon. As a result, the determination of authorship often fell prey to calligraphers, cryptologists, even charlatans.

This is no longer the case. As electronic digital computers attract more and more attention in the humanities, scholars are renewing their interest in the problem of authorship determination. In the recent past a front-page story referring to the work of the Reverend Andrew Morton, a Scots minister and mathematician, has appeared in the *New York Times* headlined: "Cleric Asserts Computer Proves Paul Wrote Only 5 of 14 Epistles." Professor James McDonough of St. Joseph's College in Philadelphia has completed his study of the *Iliad,* analyzing authorship as peripheral to his main thesis. Professor Frederick Mosteller of Harvard has earned the praise of statisticians and historians alike for his analysis of the disputed Federalist Papers. Alvar Ellegard, a Swedish scholar, has published his findings on the authorship of the *Junius Letters*.

While there is some degree of excitement in all research, authorship determination has a certain dramatic flair. The *Times* hinted at the effect of Morton's findings were they to be substantiated. Shakespeare has not yet matched wits with the computer; you can imagine the rumblings in the scholarly world if the machine spells Shakespeare M-A-R-L-O-W-E.

The work I completed in 1963 is not quite so portentous. But it did involve an authorship dispute; and it did come closer to stylistic analysis than the previous studies, which had concerned themselves chiefly with vocabulary distribution and function-word frequencies; and it was quite exciting.

My study dealt with Stephen Crane's final work, a novel called *The O'Ruddy*. No serious literary study of this book had been made, though the possibility exists that the book was a new venture for Crane into the realm of the Comic Spirit. This is a novel of about 90,000 words. Crane had completed a manuscript of about 65,000 words when he died. Considerable mystery surrounds the completion of the book, but it was finished by Robert Barr, an English novelist, three years after Crane's death. Of the original Crane manuscript only Chapter 24 is known to exist. It would seem that Crane had written to this point. Barr, however, in a letter to Crane's widow demanding additional money, claimed that "only a fourth of the book is Stephen's." Since the evidence surrounding the completion of *The O'Ruddy* points to stylistic difficulty, my analysis was made on the basis of style.[1]

It was not the purpose of this study, fortunately, to define style. Nor was it designed to examine all the stylistic elements which could be found in a prose passage. It was an attempt to discover certain aspects of style, lexical and grammatical — objective entities in that they can be counted — which would differentiate one author's writing from that of another when these composite aspects, or profiles, were compared.

A pilot study to generate techniques for such work was designed and carried out.

Since the major study was an attempt to differentiate one author from another by using stylistic variables on a scale not previously attempted, the pilot study was conservative in its choice of variables. Basic aspects of style were intermingled with items which were unusual to the writings of one or the other authors, e.g., split infinitives, non-parallel structure. In studies of authorship determination, such a procedure is quite legitimate.[2] The results of the pilot study indicated, however, that many variables which are not idiosyncratic and which appear in practically all types of writing discriminate well when their pattern of interrelationships can be used.

This finding is highly pertinent to the variables chosen for the major experiment, since only those which are basic to the writing process – in that an incipient writer should be aware of their applicability and effectiveness — were used in the final list.

The purpose of this limitation was threefold. First, style is an amorphous, all-encompassing entity which cannot be meaningfully discussed except from a specific viewpoint. Second, since such a host of elements, in combination, represent style, only certain aspects can

[1] For further discussion of the problems surrounding the completion of *The O'Ruddy* and a more complete analysis of the issues touched on in these pages, see: O'Donnell, B., *An Analysis of Prose Style to Determine Authorship: The O'Ruddy, a Novel by Stephen Crane and Robert Barr,* doctoral dissertation, Harvard 1963 (to be published). The doctoral research was in part supported by National Science Foundation grant NSF-GP-683 in cooperation with the computation centers at Harvard and M. I. T.

[2] Frederick Mosteller and David L. Wallace, "Inference in an Authorship Problem," *Journal of the American Statistical Association* 1963, 58: 275-309.

be used if they are to be handled within a statistically measurable framework. Third, this study, assuming success, took on the added significance of establishing the importance of these basic lexical and grammatical aspects of style in the study of the composing process.

The plan of using only those variables which are basic to writing style had to be further modified since an agreement among those involved in the study of composition of what is "basic" would be difficult to reach. Hence "basic" in this study meant that, in my opinion, these particular variables might be utilized in an introductory writing course. For this reason, a number of lexical and grammatical items which seemed either unique to the style of one of the authors (e.g., the unusual use of the prefix *mis-*) or too advanced stylistically to be considered basic (e.g., Crane's refined use of irony and certain other rhetorical devices) were not used, even though employing them might have enhanced prediction.

It is perhaps difficult, at first, to imagine the possible use of frequency counts of even the most meaningful variables, except for statistical purposes. A brief consideration of these counts would probably suggest their importance in the discovery of linguistic trends. Their relevance to the study of style, or more particularly to the development of style, is not so immediately evident. This relevance may become clearer if we discuss the variables from the viewpoint of their utilization.

The eighteen variables in the final listings were analyzed on the basis of their interrelationship within a paragraph. The first two, *words* and *sentences*, make up what may be called the superstructure of the paragraph. Both of course relate to length and are used by themselves as discriminators and as control factors when speaking of the frequencies of other variables. A paragraph of fifty words with five adjectives may show no difference, adjectively, from a paragraph of one hundred words with ten adjectives.

The next four variables, *clauses, dependent clauses, simple sentences,* and *past participial phrases,* might be referred to as the framework of the paragraph. In themselves they are sensitive to stylistic change. Aside from their importance from the standpoint of "what is said," involving choice for reasons of clarity, emphasis, and subordination of ideas, these variables directly relate to "how it is said," the essence of style. Ideas expressed in a series of independent clauses rather than in a series of simple sentences, or subordinated by a participial phrase rather than by a dependent clause, have a different impact on the reader even though the content is identical.

Frequency counts serve to focus attention on specific aspects of style which often, unfortunately, are taught in isolation from an actual piece of writing. Three of the major parts of speech were included among the variables: *verbs, adjectives,* and *adverbs.* The ease with which these parts of speech can be manipulated within a given context is a mark of the good writer. "But . . . but George . . . you can't . . . you can't mean that," (a) she said haltingly, (b) she stammered, (c) she said in a shocked voice — all say the same thing, but with a different nuance in meaning. The verbals, *participles* and *infinitives,* were also included for much the same reason. A consciousness of the possibility of interchange among these variables, gained by frequency counts, could then be sharpened by close inspection of the preference of one variable to another by a specific author.

Punctuation marks, necessary for clarity in any piece of writing, could be considered as part of the framework of the paragraph. Certainly periods and commas fall into this category. Other marks, such as *semi-colons* and *dashes,* have a certain flavor of their own not limited to clarity of expression. The *semi-colon* lies somewhere between the comma and the period in the degree of pause indicated. The *dash* is similar to the comma in that it requires a mental pause, but a longer one than the comma. The use of either of these variables indicates a conscious effort on the part of the writer to specifically equate or definitely separate the ideas he has expressed.

Metaphor, derived from the Greek word meaning "to transfer," may be defined: "A word that applies literally to one kind of object or idea is applied by analogy to another." It was selected, first, because the style of Stephen Crane is noted for its use; second, because of the forcefulness it adds to language; third, because it is an everyday occurrence in speech (*it's hot as hell; that comment threw him for a loss; he's a good egg*). *Color reference* was chosen for much the same reasons as metaphor (*he's yellow; he's true-blue; a scarlet woman*). It has the additional potential of introducing the young writer to basic symbols, which have no doubt been discussed with respect to his reading.

The two variables *impersonal constructions* (there is . . ., it seems . . ., etc.) and *initial conjunctions* (i.e., conjunctions appearing as the first word of a sentence) are not usually considered aspects of style that should be imitated. They were considered appropriate for this study, however, because of their value in discussing the style of writing in which they are used. *Initial conjunctions* are closely related to the variables *simple sentences* and *semi-colons.* They are usually used at the beginning of a simple sentence which may well have been attached to the preceding sentence. The sentence in which an *initial conjunction* is used often appears to be an afterthought of the writer. This pattern often is reworked in a final draft; it may be retained, however, if the after-thought effect is desired. *Impersonal constructions,* not usually found in good narrative prose, have a certain qualitative effect that usually can be better expressed by manipulation of the *verb* or *adverb* variables in a sentence. Improvement, however, does not always result from such change. It is difficult to see how the sentence "A tavern is in the town" is any improvement on "There is a tavern in the town."

Dialogue is used in this study to mean quoted expressions of one or more characters in the story which appear within a prose paragraph. That is, unquoted material precedes and follows the dialogue, all within a particular paragraph. Though not usual, this is a stylistic device in narrative writing that may be used with considerable effect: "But the new regiment was breathless with horror. 'Gawd! Saunders got crushed!" whispered the man at the youth's elbow. They shrank back and crouched as if compelled to await a flood." (S. Crane, *The Red Badge of Courage*). Since narrative writing is one of the basic forms of composition, and since this particular stylistic device is so forceful, *dialogue,* buried as it is, was included in our list of variables.

Certain of these eighteen variables were, at times, combined to form ratios, since the combination often contained meaning not attributable to a variable considered alone. In the case of *mean sentence length,* for example, not only is another meaning added to the variables *words* and *sentences* but it is a characteristic of style which has been established, in certain instances, to be a good discriminator.[3] The *verb-adjective ratio,*[4] was also included to see whether its effectiveness in differentiating between different types of writing was also applicable to discriminating authors within a specific type of writing. Certain other ratios, clause-dependent clause, sentence-verb, and clause-verbal ratios, which were used by George Hillocks in a brief study on student writing,[5] were included to see whether or not they also reflected nuances of more polished style.

This brief exposition only suggests the possibilities that these variables offer to the teaching of composition. It is an attempt, however, to establish the meaningfulness of these variables to research in this area. Further, these variables are unique in that no previous study of authorship determination has attempted to use variables which are so closely related to writing style.

Frequency counts of the chosen variables were based on sample paragraphs drawn from six novels, three written by Stephen Crane, three by Robert Barr. The novel was used since the book to be examined, *The O'Ruddy,* is a novel. Also, to insure that the fre-

[3] G. Udny Yule, "On Sentence Length as a Statistical Characteristic of Style in Prose," *Biometrika* 1939, 30: 363-390.

[4] D. P. Boder, "The Adjective-Verb Quotient," *The Psychological Record* 1940, 3: 310-343.

[5] George Hillocks, Jr., "An Analysis of Some Syntactic Patterns in Ninth Grade Themes," Euclid, Ohio, Central Junior High School (to be published).

quencies would be as representative as possible, these novels were selected on the basis that the writing covered a relatively short time span, approximately six years. Though no experimental data exist as proof, it is generally accepted that a writer's style may change over a long period of time. In a short period, however, a writer's style tends to remain stable, and any changes which do occur are not radical.

Each book was treated as a sub-population and the stratified sampling procedure recommended by Mosteller and Rulon[6] was used to obtain frequencies of the eighteen selected variables. One hundred paragraphs from each author were selected so that the subsamples were proportional to the stratum sizes. This pool of two hundred paragraphs (approximately 10,000 words) constituted the set of known samples to be used to generate and to validate the predicting system needed to determine which parts of *The O'Ruddy* were written by Crane and which parts by Barr.

The paragraph was chosen as the sampling unit for several reasons: It is a basic unit of composition and, as such, clearly pertains to the chosen variables. Also, since the number of variables is relatively high, a large number of samples must be counted to avoid a possible positive bias in the multiple correlation coefficient. Counting 18 variables in each sample unit is time-consuming; therefore, choosing paragraphs rather than pages or chapters was both expedient and procedurally correct. A final reason was that an attempt was to be made both to separate *The O'Ruddy* into *chapters* of Crane and Barr and to explore the possibility of identifying unknown writing units as small as *paragraphs* within chapters. It was felt that the weighted variables would be more sensitive for predicting purposes if they were generated from a unit of this size. The generation of these weights was accomplished by multivariate statistical procedure.

Discriminant analysis is a multivariate statistical procedure designed to answer the question of which group an individual is most like. Philip J. Rulon labels this procedure "the statistics of taxonomy" and states: "Taxonomically, doctors, lawyers, engineers, administrators, and computer technicians are looked upon as representing six species, and the distinctions between and among species are studied by multiple discriminant analysis."[7] Essentially this statistic maximizes differences among discrete classes of things such that it is possible, data permitting, to decide which group an individual most resembles. Stated another way, discriminant analysis can be used

[6] Frederick Mosteller and Phillip J. Rulon, *Principles of Statistical Inference*, chapter 8, (unpublished).

to determine the probability that a person is a member of group A or B or C, given that he is an *x*, and a *y*, and a *z*.

In terms of this study the question asked was whether a paragraph (or a chapter) was most probably written by Crane or Barr given that it contains so many verbs, adjectives, dependent clauses, etc. Considering each author's writing to be one of two species, say species C and species B, this study will attempt to use the frequencies of the selected variables to accurately determine which of these two groups a given paragraph or chapter of *The O'Ruddy* is most likely a member. For example, a section of *The O'Ruddy* which has probability greater than .5 of being a member of Group C is assumed to have been written by Crane.

A *two-group* discriminant analysis, which is the appropriate model for this research, produces results equal to multiple regression procedures using a binary criterion vector. Operationally it is irrelevant which of the two statistics is used in this research.

An advantage of using the multiple correlation form for analysis of the variables is that stepwise regression techniques can preface the attempts to establish relationships and to generate regression coefficients. The twenty-three initially selected variables are a bit unwieldy. Stepwise regression procedures were used to obtain that minimum set of variables needed for maximum predictability.

The hypothesis that at least one subset of the twenty-three selected variables was significantly related to the criterion of authorship formed the base upon which this study rested. Little reason could be found to continue this research if no relationship were found to exist between the use of the selected variables and the identity of the author. On the other hand, demonstrating that only a moderate relationship exists would make pursuit of authorship determination worthwhile.

A multiple correlation was performed on 100 sample paragraphs, 50 from each author, with 18 selected variables as predictors. The magnitude of the relationship between these predictors and the authorship criterion would indicate the force with which the initial hypothesis could be advanced. The obtained multiple correlationship coefficient was .820. Analysis of variance of R yielded an F of 9.27 which (for 18 and 81 degrees of freedom) is significant well beyond the .01 level. The multiple R^2 (.67) indicated that 67 per cent of the differences between these two authors is being accounted for by the selected variables.

To determine the amount of faith which could be placed in these

[7] Phillip J. Rulon, "Distinctions Between Discriminant and Regression Analyses and a Geometric Interpretation of the Discriminant Function," *Harvard Educational Review* 1951, 21: 80-90.

results, the standard formula for shrinkage was applied. The corrected R is .775 which remains significant. These results clearly imply not only that it is tenable to advance an hypothesis of relationship but that it is justifiable to explore the possibilities of prediction.

The task at this point, then, is to see if these generated weights are transferable to new data. The b weights (derived from the above reported multiple correlation) were applied to the raw frequencies of 100 *uncontaminated* sample paragraphs, fifty for each author.[8] Forty-four paragraphs (of fifty) of Crane and forty-three (of fifty) of Barr were correctly assigned.

These results encouraged us to continue the experiment without further refinement, at least at this point. An attempt was then made to predict on known material that was considerably larger than the paragraph. This was deemed advisable since the initial predictions in *The O'Ruddy*, to orient sections of the book, would be done on chapters. The 100 uncontaminated sample paragraphs were grouped into chunks of about 1000 words. These groups resemble chapters of the novel in that chapters are made up of several paragraphs, though in most cases the chunks are considerably smaller. The smallness of size would be a further test of the sensitivity of the variables. It is logical to assume that it is easier to predict on a larger text than on a small one because of the increase opportunity for the variables to come into play. (Mosteller,[9] Yule, and Ellegard[10] all consciously attempt to avoid predicting small texts. Ellegard goes further in saying that 1000 words is the minimum text size on which accurate prediction can be made.)

The weights of the 18 variables were applied to raw frequencies of the variables in the 1000-word chunks. The results establish the transferability of these weights. The fact that all groups were assigned accurately — and substantially so — demonstrates the applicability of the generated system to relatively large units of several paragraphs. Attempts to predict chapters of *The O'Ruddy* could proceed with a certain confidence that assignments would be accurate.

The results of the application of the weighted variables to the chapters of *The O'Ruddy* were both dramatic and forceful. Chapters 1 through 24 were assigned to Crane; Chapter 25 emerged as a transitional chapter; Chapters 26 to 33 were strongly Barr's work.

[8] The term "uncontaminated" is applied to data which has not been used to generate the predicting weights; hence, an uncontaminated *known* sample may be treated as an unknown as a test of the procedures.

[9] Mosteller and Wallace, *loc. cit.*

[10] Alvar A. Ellegard, *A Statistical Method for Determining Authorship*, Stockholm, 1962.

The results of prediction of the individual paragraphs within the chapters were just as dramatic; they lend themselves, however, to further evaluation, possibly of a more literary nature. In Chapter 1, for example, two of the twenty-four paragraphs were assigned to Barr. Two possibilities exist here: Crane wrote them and they were mis-assigned (we do admit the possibility of some error); or Barr inserted them in this chapter for a specific reason. A literary analysis of the mis-assigned paragraphs would shed light on this. The results for our purposes, however, were quite satisfactory.

The point has been often made in this discussion that the stylistic variables used in this experiment are in themselves meaningful to those interested in developing student writing. The variables were chosen particularly with this relevance in mind. Hence, this study might offer to those researchers in composition an adaptable method for discovering lexical and grammatical patterns in student writing. If the pattern of a student's writing can be concretized, the task of spotting weaknesses and suggesting changes to correct them will be considerably lightened for both the teacher and the student. This analytic method could well be applied to examining "good" writing — the writing of accepted authors. The specific stylistic devices that these writers use would be identified and underscored by counts and correlations. Thus, for the first time perhaps, a student would understand what is meant by the Hemingway style or the Faulkner style.

Research in the field of composition is only beginning to develop. These grammatical and lexical variables, which were chosen for this study, could easily be overlooked because of their familiarity. It was the way in which these variables were used by Crane and Barr, however, that differentiated between their styles; therefore, the results of this study serve to highlight the importance of the variables in the study of the composing process.