Statistical and Linguistic Strategies in the Computer Grading of Essays

Ellis B. Page
University of Connecticut
Storrs, Conn., U.S.A.

Essay tests are used in the schools and colleges of all nations, and in major testing programs of national and even international size. Potentially, such essay tests are an important applied field for computational linguistics, and should eventually provide focus for much work. Yet in the past, little direct attention has been paid to such grading, although there are ways to begin investigation which would not necessarily require much linguistic knowledge beyond that now available.

Beginning in December of 1964, Project Essay Grade (PEG), at the University of Connecticut, has investigated the computer analysis and evaluation of student writing. In February, 1965, the project was given pilot funding by the College Entrance Examination Board of New York City, and in June, 1966, the United States Office of Education gave it much larger support. Through this period of preliminary investigation, certain problems have become much better understood (Daigon, 1966; Page, 1966, 1967). This paper discusses these problems, relates certain major findings to date, and outlines apparently promising avenues for future work by linguists, computer scientists, psychologists, and educators.

Background

It is useful to conceptualize the field of essay grading in two dimensions, as represented in Figure 1.

Figure 1

Possible Dimensions of Essay Grading

	I Content	II Style
A. Rating Simulation	I (A)	II (A)
B. Master Analysis	I (B)	II (B)

Any serious effort to grade essays must obviously face problems of "content" as in Column I, and of "style" as in Column II. Yet it is obvious that these columns are not mutually exclusive. Similarly, the rows are not mutually exclusive either, but their general meaning must be mastered to understand the work to date and the problems of the field. The first row refers to the

simulation of the human judgment, without great concern about the way this judgment was produced. The second row refers to the accurate, deep, "true" analysis of the essay.

We have coined two terms to describe this difference. Since the top row is concerned with approximation, we speak of the computer-variables employed as proxes. Since the bottom row is concerned with the true "intrinsic" variables of interest, we speak of such variables as trins. A trin, then, is a variable of intrinsic interest to the human judge, for example, "aptness of word choice". Usually a trin is not directly measurable by present computer strategies. And a prox is any variable measured by the computer, as an approximation (or correlate) of some trin, for example, proportion of uncommon words used by a student (where common words are discovered by a list look-up procedure in computer memory).

In the early part of our investigations, we concentrated on the right column and top row of Figure 1, looking for actuarial strategies, seeking out those proxes which would be of most immediate use in the simulation of the final human product, the ratings of stylistic factors.

For the first attempts, we evolved a general research design, which we have more or less followed to date:

(1) Samples of essays were judged by a number of independent experts. For our first trial 272 essays, written by students in Grades 8 to 12 in an American high school, and judged by at least four independent teachers. These judgments of overall quality formed the trins.

(2) Hypotheses were generated about the variables which might be associated with these judgments. If measurable by computer, and feasible to program within the logistics of the study, these computer variables became the proxes of the study.

(3) Computer routines were written to measure these proxes in the essays. These were written in FORTRAN IV, for the IBM 7040 computer, and are highly modular and mnemonic programs, fairly well documented and available to computational linguists interested in using them or adapting them.

(4) Essays were prepared for computer input. In the present stage of data processing, this means that they were typed by clerical workers on an ordinary keypunch. They were punched into cards, and these cards served as input for the next stage.

(5) The essays were passed through the computer, under the control of the program which collected data about the proxes. The output was as appears in Figure 2.

Figure 2

PEG-IA Output

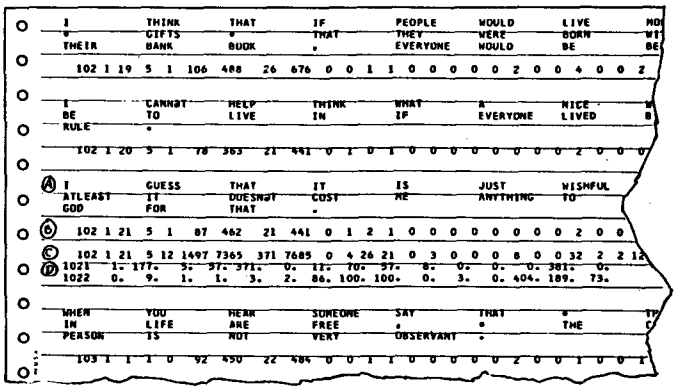

Figure 2 shows a piece of output from PEG-IA. Line A shows the way a sentence from the student essay is re-written in 12-character double-precision computer "words" and stored in memory. Line B shows the summary of data for that sentence just analyzed. The first number is the essay identification. The other numbers of Line B are some counts from that sentence. Line C shows a summary of these counts, across sentences, for this whole essay. And Line D are these measures transformed in a number of simple ways, and ready for input into the final analysis.

(6) These scores were then analyzed for their multivariate relationship to the human ratings, were weighted appropriately, and were used to maximize the prediction of the expert human ratings. This was all done by use of standard multiple-regression programs.

The first analyses produced results as shown in Table 1. Here it is possible to read the list of proxes (Col. A), and their correlation, after transformation, with the human judgments of overall quality (Col. B). Col. C shows their contribution to the total multiple regression, and Col. D indicates the test-retest reliability of the proxes themselves, as discovered from two essays written by the same students, with about a month between writings.

Table 1*

Variables Used in Project Essay Grade I-A for a Criterion of Overall Quality

A. *Proxes*	B. *Corr. with Criterion*	C. *Beta wts.*	D. *Test-Ret. Rel. (Two essays)*
1. Title present	.04	.09	.05
2. Av. sentence length	.04	−.13	.63
3. Number of paragraphs	.06	−.11	.42
4. Subject-verb openings	−.16	−.01	.20
5. Length of essay in words	.32	.32	.55
6. Number of parentheses	.04	−.01	.21
7. Number of apostrophes	−.23	−.06	.42
8. Number of commas	.34	.09	.61
9. Number of periods	−.05	−.05	.57
10. Number of underlined words	.01	.00	.22
11. Number of dashes	.22	.10	.44
12. No. colons	.02	−.03	.29
13. No. semicolons	.08	.06	.32
14. No. quotation marks	.11	.04	.27
15. No. exclamation marks	−.05	.09	.20
16. No. question marks	−.14	.01	.29
17. No. prepositions	.25	.10	.27
18. No. connective words	.18	−.02	.24
19. No. spelling errors	−.21	−.13	.23
20. No. relative pronouns	.11	.11	.17
21. No. subordinating conjs.	−.12	.06	.18
22. No. common words on Dale	−.48	−.07	.65
23. No. sents. end punc. pres.	−.01	−.08	.14
24. No. declar. sents. type A	.12	.14	.34
25. No. declar. sents. type B	.02	.02	.09
26. No. hyphens	.18	.07	.20
27. No. slashes	−.07	−.02	−.02
28. Aver. word length in ltrs.	.51	.12	.62
29. Stan. dev. of word length	.53	.30	.61
30. Stan. dev. of sent. length	−.07	.03	.48

*Number of students judged was 272. Multiple R against human criterion (four judges) was .71 for both Essay C and Essay D (D data shown here). F-ratios for Multiple R were highly significant.

The overall accuracy of this beginning strategy was startling. The proxes achieved a multiple-correlation of .71 for the first set of essays analyzed and, by chance, achieved the identical coefficient for the second set. Furthermore, the beta weightings from one set of essays did well in predicting the human judgments for the second set of essays written by the same youngsters. All in all, the computer did a respectable, "human-expert" job in grading essays, as is visible in Table 2.

Table 2

Which One is the Computer?

Below is the intercorrelation matric generated by the cross-validation of PEG I

	Judges				
	A	B	C	D	E
A		51	51	44	57
B	51		53	56	61
C	51	53		48	49
D	44	56	48		59
E	57	61	49	59	

Here we see the results of a cross-validation. These are correlations between judgments of 138 essays done by five "judges," four of them human and one of them the computer. The computer judgments were the grades given by the regression weightings based on 138 other essays by other students. This cross-validation, then, is very conservative. Yet, from a practical point of view, the five judges are indistinguishable from one another.

However useful such an overall rating might be, we of course still wished greater detail in our analysis. We therefore broadened the analysis to five traits believed important in essays, adapted partly from those of Paul Diederich. They may be summarized as: ideas, organization, style, mechanics, and creativity. We had a particular interest in creativity, since some critics from the beginning have believed that the computer must founder on this kind of measure. "You might grade mechanics all right," someone will say, "but what about originality? What about the fellow who is really different? The machine can't handle him!"

Therefore, in 1966 we called together a group of 32 highly qualified English teachers from the schools of Connecticut to see how they would handle creativity and these other traits. Each of 256 essays was rated on a five-point scale on each of these five important traits, by eight such expert judges, each acting independently of any other judge. The teacher ratings were then analyzed, and it was found that the essay and the trait contributed significant variances, as did the trait-by-essay interaction, (perhaps the clearest demonstration of the ipsative profile). To investigate each of these five trait ratings, the same 30 proxes were again employed, with the results to be seen in Table 3.

Table 3

Computer Simulation of Human Judgments
For Five Essay Traits
(30 predictors, 256 cases)

A. Essay Traits	B. Hum.-Gp. Reliab.	C. Mult. R	D. Shrunk. Mult. R	E. Corr. (Atten.)
I. Ideas or Content	.75	.72	.68	.78
II. Organization	.75	.62	.55	.64
III. Style	.79	.73	.69	.77
IV. Mechanics	.85	.69	.64	.69
V. Creativity	.72	.71	.66	.78

Note:

Col. B represents the reliability of the human judgments of each trait, based upon the sum of eight independent ratings, August 1966.

Col. C represents the multiple-regression coefficients found in predicting the pooled human ratings with 30 independent proxes found in the essays by the computer program of PEG-IA.

Col. D presents these same coefficients, shrunken to eliminate capitalization on chance from the number of predictor variables (cf. McNemar, 1962, p. 184)

Col. E presents these coefficients, both shrunken and corrected for the unreliability of the human groups (cf. McNemar, 1962, p. 153.)

In our rapidly growing knowledge, Table 3 may temporarily say the most to us about the computer analysis of important essay traits. Column A of course gives the titles of the five traits (more complete descriptions of the rating instructions may be supplied on request). Column B shows the rather low reliability of the group of eight human judges, computed by analysis of variance.

Here in Column B "creativity" is less reliably judged by these experts than are the other traits, even when eight judgments are pooled. And mechanics may be the most reliably graded of these five traits. Surely, then, humans seemed to have a harder time with creativity than with mechanics.

What of the computer? Column C shows the raw multiple correlations of the proxes with these rather unreliable group judgments. These were the coefficients produced by the standard regression program run by Dieter Paulus and myself. Column D simply shows the same coefficients after the necessary shrinking to avoid the capitalization on chance which is inherent with multiple predictors. Finally, in order for a fair comparison to be made among the traits, the criterion's unreliability should be taken into account, as in Column E. Here such difficult variables as creativity and organization no longer seem to suffer; the computer's difficulty is apparently in the criterion itself, and is therefore attributable to human limitations, rather than to machine or program limitations. Column E, then, exhibits what might be the expectable cross-validation from a similar set of essays, if predicting a perfectly reliable set of human judgments.

Current and Projected Problems

Of course, all this is a temporary reading taken in the middle of the research stream. Our investigators have also gone on with other strategies. Donald Marcotte (1967) has developed a phrase analyzer, and has discovered that clichés, as usually listed, were largely irrelevant to the judgment of such essays. Dieter Paulus (1967a) has studied the curvilinearity of proxes, and concluded that much elaborate statistical optimization may be a waste of time, and that the most major improvements should probably be made in other ways. He also has studied feedback to the student writer, using an on-line time-sharing console (Paulus, 1967b), as has also Michael Zieky. Another researcher, Jack H. Hiller (1967), has investigated quasi-psychological dimensions (including opinionation and vagueness) as predictors of the human judgments. Using techniques familiar from automatic content analysis (cf. Stone et al, 1966), he constructed lists of words and phrases to define the variables of psychological interest, and found these negatively correlated, as he predicted, with writing quality. And, in May, 1967, a sizeable improvement was made in the statistical accuracy, increasing the multiple-regression coefficient from about .71 to about .77, and improving the variance accounted for by around 20%. In other words, the newest programs apparently do better than the individual, expert English teacher.

The early strategies, then, have provided fertile ground for statistical investigation of essay grading, especially in the actuarial simulation of rating of style. But what of the deeper dimensions of stylistic analysis, and what of subject-matter content, as in essay questions in history, philosophy, or science?

Possible contributory linguistic strategies have been under more intensive study in recent months, with the advice and help of Susumu Kuno (1964), Stanley Petrick (Keyser and Petrick, 1967), John Olney (Olney and Londe, 1966; also see Harris, 1952) and others. (Of course these workers are not responsible for errors or misconceptions in the present paper.) Anticipated future strategies are currently summarized in Table 4. This table is based partly on work already accomplished in Project Essay Grade, partly on suggested minor adaptations of systems already working for others, and partly on projected programs which are not yet apparently operative in any system, but which do not seem impossibly difficult at the efficiency desired.

Table 4

Project Essay Grade Hypothetical Complete Essay Grader
1. INPUT and PUNCH. Handwritten or typewritten or other raw response of the writer is converted for computer input.
2. SNTORG. Creates arrays of words and sentences as found in prose. This is just as performed in PEG-I.
3. DICT. Assignment of available syntactic roles to each word. This is currently done by many programs, but needs an expanded dictionary, and ambiguity resolver. At the same time, the semantic information will be stored in the workspace for reference of other parts of program. Availability of the tape-written Random House Dictionary (Unabridged) has been promised.
4. PARS. A modified Kuno (1964) program seems most promising, and is currently being programmed for both the 7094 and the 360 by workers at IBM. Alterations will be necessary to accept well-formed substrings.
5. REFER. This is intended to identify and encode the most likely referents of pronouns and other anaphoric expressions. (Cf. Olney and Londe, 1966). This process must employ both syntactic features and semantic information from DICT.

(Continued)

Table 4 (Continued)

6. KERNEL and STRUC. From the rewritten string output of (5), KERNEL would establish a set of elementary propositions, and STRUC would encode the relationships among these elements. This step would retain all the information of an essay in simplest possible units, yet would retain additional information about emphasis, subordination, causal relation, etc., among these units.

7. EQUIV. The elementary units would be augmented by the semantic information in DICT. To each word would be assigned a cluster of permissible synonyms, with weightings of semantic distance. This permits an analysis of redundance and emphasis in the essay, and permits a comparison of the content of the student essay with that of the key or master essay.

8. STYLE. Descriptions of the surface structure characteristics of the essay: parts of speech, organization of themes, types and varieties of sentence structure, grammatical depths, tightness of reference, etc; information about grammatical errors and strengths.

9. CONTNT. Comparison of the agreement of student and master essay, through measure of kernel hits and struc hits, these weighted by semantic distance of language chosen.

10. SCOR. Multivariate prediction of appropriate profile for the immediate purpose.

The limitations of space will permit only a few comments on this table, which may be seen as representing a hypothetical, ideal essay grader. For large grading systems, over established substantive content, it would be possible, for the key or master essay, to edit by hand the output from certain routines (especially REFER and STRUC). Of course, four of the most important routines listed in Table 4 are far from perfected in any existing programs. Ideally, they would assume better solutions to certain major, stubborn problems in computational linguistics.

Indeed, the steps in this hypothetical essay grader are close to the heart of the most persistent and troublesome problems in linguistics. Is it necessary that sentences be syntactically analyzed before mapping into deep structure? What is the proper role of semantics in such deep structure? How can the outside knowledge of the reader be incorporated into the machine analysis? (For some discussion of this problem, see Quillian, 1966). In general, how may we incorporate some of the intuitive richness which the literate human brings to his reading?

It is not expected that workers in essay grading will suddenly resolve all such questions. They may be recognized as those which so trouble linguists as to contribute to the recent official pessimism, in the United States, about the future of mechanical translation. After 15 years of effort, mechanical translation is still regarded as disappointing in quality, and virtually no sustained output of any machine program would be ordinarily mistaken for the work of a professional human translator.

On the other hand, the earliest attempts at essay grading by computer have, in a very limited way, leaped ahead of machine translation. And if the expert human ratings of high school essays may be regarded as an acceptable goal, then the machine program appears to have reached such a goal already. For that matter, improved performance, even superior to that of the individual human expert, appears to be immediately practicable as well.

The explanation of this advantage, of course, is that the problem of essay grading as attacked in the current work is much easier than the problem of machine translation. In translation, every nuance of the input string should be accounted for in the output string. In essay grading, only a certain portion of the input text needs to be accounted for, and the output does not depend on the existence of any large language-generating system. High quality machine translation apparently demands a fair portion of the total language-manipulating capability of the human, but essay grading may use only a fraction of it, and may process language in ways quite different from that of the human being. For example, our present programs have to date largely ignored order and sequence in the essays, although to the human the order of words is, of course, of crucial and unceasing importance.

Since essay grading can work with such fractional information, then, why pursue the deeper analysis of Table 4? Clearly, the purpose is not entirely the same as it would be for the usual linguist. At any discrete

time in research, what is sought is not necessarily the perfect humanoid behavior, but rather those portions of that behavior which, given any current state of the art, will contribute optimally to efficient and practicable improvements in output. Indeed, regardless of the eventual perfection of deep linguistic behavior, for any specific application to essay grading, at any one moment, large portions of such available behavior may be irrelevant, just as it seems that ordinary human language processing does not usually call for our full linguistic effort.

Yet we regard it as eventually important to be able to perform these various kinds of advanced machine analysis when required. Therefore, the eventual uses of the ideal essay analyzer may require analytic capability as deep as may be imagined. Writing out suitable comments for the student, for example, will in some cases tax any system which may be foreseen.

Even approximate solutions to these problems, however, though unsatisfactory for certain scientific purposes, could make important contributions to the educational description and evaluation of essays. For such evaluation is itself probabilistic, limited by imperfect asymptotes of writer consistency and rater agreement. And such evaluation therefore does not require, to be practicable and satisfactory, the same deterministic perfection which has continued to elude and frustrate researchers in mechanical translation. There is a fundamental difference in goals, which must be realized. As has been demonstrated here, the output from much cruder statistical programs has already reached a quality not too remote from usefulness. The more advanced strategies currently seem, at least to the present workers, bright with promise.

Daigon, Arthur. Computer Grading of English Composition. The English Journal, January, 1966, 46-52.

Harris, Z. S. Discourse analysis. Language, 1952, 8 (4), 474-493.

Hiller, Jack H., Page, E. B., and Marcotte, D. R. A Computer Search for Traits of Opinionation, Vagueness, and Specificity-Distinction in Student Essays. Paper read at the Annual Meeting of the American Psychological Association, Washington, D.C., September 2, 1967.

Keyser, S. J., and Petrick, S. R. Syntactic Analysis, 1966. (In press in a forthcoming book.)

Kuno, Susumu. Some characteristics of the Multiple-Path Syntactic Analyzer. Language Data Processing, Cambridge: Harvard Computation Laboratory, 1964. C6, 1-8.

Marcotte, Donald. The Computer Analysis of Cliché Behavior in Student Writing. Paper read at the Annual Meeting of the American Educational Research Association, New York, February 18, 1967.

McNemar, Quinn. Psychological Statistics, 3rd ed. New York: Wiley, 1962.

Olney, John and Londe, D. A research plan for investigating English discourse structure with particular attention to anaphoric relationships. Tech Memo mm-(L)-3256. Santa Monica, California: System Development Corporation. November 22, 1966. 17 p.

Page, Ellis B. The Imminence of Grading Essays by Computer. Phi Delta Kappan, January, 1966, 238-243.

Page, Ellis B. Grading Essays by Computer: Progress Report. Proceedings of the 1966 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1967. Pp. 87-100.

Paulus, Dieter. Problems of Nonlinearity in Grading Essays. Paper read at the Annual Meeting of the American Educational Research Association, New York, February 16, 1967a.

Paulus, Dieter. Feedback in Project Essay Grade. Paper read at the Annual Meeting of the American Psychological Association, Washington, D.C., September 2, 1967b.

Quillian, M. Ross. Semantic Memory. Cambridge, Mass.: Bolt Beranek and Newman, 1966.

References (Continued)

Stone, Philip J., Dunphey, Dexter C., Smith, Marshall S., and Ogilvie, Daniel M. The General Inquirer: A Computer Approach to Content Analysis. Cambridge: M.I.T. Press, 1966. Pp. 651.

Woods, William A. Semantics for a Question-Answering System. Paper read at the Annual Meeting of the Association for Machine Translation and Computational Linguistics. Atlantic City, N.J. April 21, 1967.