Winged words: varieties of computer applications to literature

by LOUIS T. MILIC
Columbia University
New York, New York

On August 6, 1961, an item appeared in the *New York Times* describing a rather unusual project that a young classical scholar named James T. McDonough had been actively pursuing. According to the account, for the previous four years McDonough had been reducing the Iliad of Homer to patterns representing the meter of the Greek epic, key-punching these patterns on a large number of punch cards and running them on an IBM 650 in the hope of discovering whether the uniformity of the patterns showed that the poem had been written by a single author, a question about which there had been a great deal of argument during the past century. McDonough hoped to earn a Ph.D. at Columbia with this work. Whether he succeeded or not, he had made a landmark in being one of the first who used computers in the solution of a literary problem. Since that beginning, others have availed themselves of electronic aid. As it is now ten years since the beginning of the relationship between computers and literature, it may be time to survey some of the results, to consider what has been achieved and what the prospects are.

In talking about literary computation, I want to confine myself to activities which are strictly literary. There are a number of related activities which border on literature but which are really tangential or preliminary to literary study. Among these I include machine translation, the making of concordances, attribution study, editing and bibliography, as well as most kinds of linguistic study. The reasons are relatively simple.

Machine translation arose out of a suggestion made in 1947 that machines might help to deal with the large bulk of foreign-language materials, especially in Russian, that government agencies and scientists found it necessary to go through in order to keep up with current developments. At first, owing to certain oversimplified ideas about the structure of languages, this seemed like a reasonable task to ask a computer to perform. Naturally the government was interested in furthering this research and invested substantial amounts in hardware and programs. The linguists were excited by the challenge and devised ever newer grammars to deal with the binary nature of the machines. Gradually, the hope of success dimmed as a realization of the incredible complexity of natural language gradually emerged. Not machine limitations but inadequate knowledge about the processes of the human mind and the nature of language finally doomed the machine translation project.

A year ago, the Automatic Language Processing Advisory Committee of the National Research Council issued a report in which it was concluded that, although machine-aided translation might be deserving of further support, pure machine translation could no longer be considered a practical possibility and ought not to receive further financing. The difficulties that machine translation programs have in distinguishing between literal and metaphoric uses of language and in dealing with idioms and problems of context are notorious. They can be illustrated by two examples, probably apocryphal. The expression "out of sight, out of mind" was supposed to be translated into Russian, where it became "invisible idiot." The phrase "the spirit is willing but the flesh is weak" returned from the Chinese analyzer as "whisky O.K., meat no good."

Whether these are true or not, they suggest the superiority of the human mind in dealing easily with levels of literalness. The twenty years and the millions spent on the effort to develop translation programs were not wasted on a mere effort to inflate the human ego. Though little translation actually took place, a great deal was learned about the nature of language. It may be claimed that postwar linguistics was revolutionized by the discoveries of machine translators and their auxiliaries, the mechano-linguists.

The field of computational linguistics is very active now and a great deal of valuable research is taking place in it which will ultimately be of use in translation and even in literary analysis. Work in automatic syntactic analysis, sentence generation, semantics has implications

for all kinds of word-connected activity. It is, however, remote from the immediate concern of the literary scholar and is not properly included in literary computation. All these linguistic projects are related to the scientific study of language, which today has the status of a behavioral science. The study of language as the medium of literature is quite a different realm, however much it may overlap that of linguistics, because its main concern is not with the characteristics of the code itself but with the individual and aesthetic use of the resources of language.

For that reason another whole class of studies may be excluded from consideration here, although they seem to reflect a much nearer concern with literature. I refer to concordances, dictionaries, glossaries, indices verborum, word-lists and bibliographies. Computers have vastly facilitated the compilation of these tools, and these tools do have something to do with the study of literature, but the use of computers in literary research is something different from the mere construction of lexical works of reference, which have existed in one form or another for centuries and which, when completed, are not essentially different from their handmade predecessors. The essential use of computers in literary research should diverge both quantitatively and qualitatively from the conversion of manual to electronic processing of data. The computer has made possible the processing of information in such quantities that no man's lifetime or energy could previously have contained it.

The mere amount of computer processing is a kind of innovation that we owe to technology. The class of studies that best fits this description of quantitative innovation is that of attribution, which has developed considerable activity since the advent of computers. To attribute an anonymous or uncertain work to its author requires processing a substantial corpus of text for each possible author and comparing its features with those of the work in question. Previously such attributions were made impressionistically on the basis of intuitively-perceived similarities or differences which could only be vaguely described: "This poem or this essay *sounds* like the work of Pope or Shelley or Ruskin." The inherent characteristics of the computer have necessitated the formalization of aspects of style for electronic processing. The predominance of short sentences, or of certain types of function words, the presence of certain grammatical constructions or favored lexical items, intervals between successive conjunctions, statistical properties of sentences or word-length distribution are examples of formal features of style.

The resultant combination—large corpora of text and cmpirical features of description—has made possible the identification of disputed works in ways that could not previously be imagined. Those who conducted the attribution studies on *The Federalist* papers, the *Letters of Junius* and the Epistles of St. Paul dealt in millions of words and have lived to tell about it. In the process of providing descriptions of the text explicit enough for the machine, they have added to our knowledge of these texts. Naturally, their results have not found favor everywhere, but they are on sound ground statistically and they have in fact merely ratified prevailing opinions in all three cases. Doubtless many more such studies will be undertaken now until the supply of disputed works runs out. One may look forward if he wishes to a definitive settling of the Bacon-Shakespeare-Marlowe contest.

Despite their usefulness to students of late eighteenth century political writing and of New Testament Greek, these studies are not essentially literary either. Their main interest is historical. They answer the question: "*Who* wrote this?" The literary information produced is merely a byproduct of the investigation. To be sure, a succession of attribution studies would provide extremely valuable information—information of which we have but the outlines at present—about the historical development of the English language. Such information, however, is linguistic rather than literary in nature. It is the background for stylistic studies but it is not itself literary. It is related to but not a fundamental part of the basic literary questions, which underlie the vast mass of literary scholarship.

If we now look at what has occupied scholars during the first decade of literary computation, we may be able to say whether they have been concerned with literature or with the preliminaries. Turning first to published work, we have some large projects resulting from the cooperation of a number of individuals and institutions. The Cornell Concordances, jointly fathered by Cornell and IBM, now cover Matthew Arnold, Emily Dickinson, and William Butler Yeats. The same group has plans for a number of additional works including most of the English poets whose works have not yet been so favored. A French group at Besancon has been conducting studies of the French vocabulary and making word-indexes of French poets and playwrights, publishing their results in two periodicals of their own. One of their separate publications is a concordance of Baudelaire which came out in the same year as one made by an individual scholar in this country. There is also a concordance to the Revised Standard Version of the Bible, which was published right at the beginning of our decade and a number covering medieval works in English and German which have just come out. It seems clear that for many scholars, using a computer has meant making a concordance.

Another large project was the attempt to solve the attribution problem in *The Federalist Papers*. Alexander

Hamilton and the editor of the *Papers,* James Madison, had long been considered in contention for the honor of having written a certain number of these pieces. Many historians were inclined to give them all to Madison despite the circumstantial evidence for Hamilton's claims to authorship. Two statisticians, one at Harvard and one at Chicago, decided to test the value of the Bayes theorem by applying it to this problem. With the help of a corps of assistants, two computers and a variety of government grants, Mosteller and Wallace concluded, as the scholars had done, that Madison had written them all. Because their concern was in statistics rather than in literature, their results do not have much interest for literary scholars. The work of a Swedish student of English literature, Alvar Ellegard, on a similar problem, the authorship of the *Junius* Letters, has been more interesting because of the information about the language of this period that he turned up. His conclusion about the authorship of the Letters coincided with prevailing opinion.

The researches of the Rev. Mr. Andrew Q. Morton of Scotland and his statistical colleagues is in a slightly different category. They have tried to distinguish between the various Epistles of St. Paul, the genuine and the spurious. Partly because of the manner in which his claims were presented and partly because of some sense among the public that the final sanctuary had been invaded by the machine, Mr. Morton has called down on his head the anger of a great number of people, including even some of the members of his cloth who are themselves using computers. Morton's results have not been fully made public, but he seems to have also found himself in agreement with previous, manually-assisted, scholars in his field. The criteria he used are not unlike those applied to *Junius* and the *Federalist*—frequency and distribution of function words in the text —but the text he uses is necessarily less reliable than theirs. After all, original copies of the eighteenth-century journals still survive but the text of St. Paul is in altogether a different state. The controversy continues to give off energy.

Questions which may be considered editorial were tackled by two scholars who used a similar technique in widely separated places. Both took advantage of the computer's ability to make a great many precise comparisons in trying to decide by means of spelling which text of Dryden or of Shakespeare had greater authority. In a sense the procedure is like that of the concordance maker with one important difference. A concordance program can only with difficulty be adjusted to recognize spelling variants of the same word. The studies just mentioned took advantage of this limitation in discovering spelling variation.

Projects even more remote from strictly literary work have been done and include a bibliographic index to the whole run of a Spanish literary journal, a million-word corpus of modern American English—this latter not published but stored on tape and available for consultation—and some collation and editing procedures that are of interest only for technical reasons.

Some very ambitious pilot studies have emanated from the workshops of Mrs. Sally Sedelow, now of Chapel Hill. Her interest, like my own, is in computational stylistics, the study of idiosyncratic patterns in individual writing. One of her programs converts specified verbal characteristics into graphic equivalents for easier comparison. Thus each noun and verb in a text could be indicated by a particular symbol, all other words being represented by zeroes. The noun-verb distribution would then be clearly visible and could then be evaluated. The trick of course is to think of the right things to look at, things that will tell us something about the text. The other program is more conventional, in the sense that it resembles a technique already in use for some time in the social sciences and named Content Analysis. The General Inquirer system, only recently applied to literary problems is a well-known example of a computer implementation of this technique. In essence, it consists of a thesaurus of themes and categories. If a text contains a sufficient selection of terms from a given category, it is concluded that the writer was concerned with that theme. Thus Mrs. Sedelow concludes from the number of words about lunacy (*mad, madly, madness, insane, disease*) in the first act of *Hamlet* that Shakespeare had this in mind when he wrote the play. Doubtless more esoteric conclusions can be reached by studying word-clusters and word-associations. At any rate this approach has the virtue of attacking the semantic component of language, which has been a great problem to all literary users of computers.

Not to overlook present company, I should also mention Professor Raben's well-known study of the influence of Milton on Shelley. In trying to pinpoint this debt, he tried to find how often in any sentence, Shelley used Milton's actual words. Contrary to his most optimistic estimates, he found an amazing number of such uses, clearly demonstrating the extent to which the later poet had incorporated into his mind the words of his predecessor. As might also have been expected, the handling of poems running into 200,000 words in the aggregate caused a certain number of space problems in the computer itself.

My own study of the style of Jonathan Swift, part of which was done on an IBM 1620, may perhaps be properly added to the end of this list, at least because it was only published this year though completed in 1963. My concern was to discover the individual features of this writer's style and to draw some literary conclusions

from this. After programming, the main technical problem I faced was the large amount of time that my list-processing procedures were using up.

As can be seen from this list, all but a few of these results of applying computers to literature have produced data sure to be useful to literary scholars—works preliminary to literary study—but are not themselves literary studies for the most part.

If we move now to work in progress as it is listed in the May issue of *Computers and the Humanities,* we find a vastly increased amount of activity. There are 120 projects listed under "Literature," though some scholars are responsible for more than one. Under examination, these break down into the following components. Predictably enough, the largest class (53) consists of concordances, dictionaries, word-lists, indexes, and catalogues of lexical items. The second largest category (25) includes various kinds of linguistic studies, programs for analyzing the linguistic characteristics of languages rather than of authors. There are seven bibliographical projects and six concerned with editing, collating, formatting and text history. Another six are devoted to various aspects of content and semantic analysis and the discovery of keywords. Five are attribution studies and another five are studies of meter and rhyme. Four are in machine translation. Of the remaining nine, two represent attempts to work up programs to serve literary scholars and are therefore really projects in information processing. This leaves seven which can be classified as strictly literary.

The descriptions provided are not full enough to permit complete understanding but it is possible to hazard some guesses as to what these projects may attempt. Two are studies of individual writers, one on a psychological basis involving word or image clusters, the other through his syntax. There is an attempt to determine whether a sonnet style exists. A comparison between a book of proverbs and a play is intended to show the reliance of the dramatist on proverbial sayings. There is a census of the roles of actors during a certain period to determine the nature of their specialization. And there is a study of the relation of grammatical deviation to mental disturbance, a matter of some interest considering how many poets have been or have been considered crazy. Except for the emphasis on linguistics, the distribution is similar to the earlier one.

Anyone who was not aware of the computer implementation of these projects and compared them to those recorded in such a Bibliography as the one published annually by the Modern Language Association might reach the conclusion that a revolution in the study of literature had taken place. Nearly half of the projects devoted to making reference-lists, nearly a quarter to linguistics! To be sure, the two samples differ considerably in size. The current issue of the PMLA Bibliography, recording almost exclusively items published in 1966, contains more than 20,000 entries covering work on all the major European languages since the Middle Ages. In it there is a small sub-sub-section on Computer-Assisted Literary Research, which contains some forty items, some of them merely general or popular explanations. At most this activity represents a very small fraction of the admittedly excessive total: one-fifth of one per cent, or one literary scholar in 500 is working with computers.

Because there is no classification of the items by type in the *PMLA* Bibliography, it would be very time-consuming to draw up a table, similar to the one just presented, for the efforts of traditional scholars. A casual examination of a random 120 items reveals, however, a predominance of historical studies of texts and documents, related social and political investigations, explications and criticisms of individual works, as well as some stylistic and linguistic studies, probably based, as is generally the case, on inadequate data. Without question, a number of all these studies could have benefited considerably from the data-gathering and data processing power of computers. In fact, it is probable that some studies are of doubtful validity because of the unrepresentative nature of their database. Traditional literary scholarship is notorious for extrapolations that go vastly beyond the data and even for conclusions reached without primary data of any sort.

What this suggests about the relationship between traditional and computer-assisted literary research is that both kinds of scholars seem to be pursuing the same ends but that what I may perhaps call the "modern" scholar has in the main limited his scholarship to certain kinds of preliminary work which is the basis for conclusions of a more far-ranging character. Concordances and the like permit studies of works and authors to be more soundly based. Attribution studies enable the critic to feel more positive about the canon of an author's work. But all these studies ultimately serve the same master. To be meaningful they must stand in a certain relation to the basic critical questions which determine the nature of any art.

What are these literary questions to which such deference must be paid? They are all primarily founded on the aesthetic aspect of human activity, the third member of the Platonic trinity of the good, the true and the beautiful. More specifically, literary criticism and scholarship must concern themselves with distinguishing between the aesthetic and the everyday, good literature and bad, poetry and mere verse. In so

doing, the scholar must give his attention to the nature of the aesthetic effect, the creative activity of the writer as opposed to the merely routine aspect of communication. This question was, until recently, unique with literature because its practitioners use the same language in writing odes and sonnets as is used in the daily newspaper, the freshman theme and manuals of instruction for computers. Pop art (the conversion of tomato soup cans and giant hamburgers into the substance of art), the underground film (the 8-hour showing of a man sleeping), and certain tendencies in music (the bizarre use of musical silences and ugly sounds), have, however, eroded the uniqueness of this feature peculiar to literature. These other arts have now been compelled to take a stand on the basic aesthetic question "What is art?" before being able to arrive at the next one, "What is good art?"

For literary students, the basic question remains "What is literature?" In the process of trying to answer it, the scholar finds himself dealing with a variety of subordinate questions, the answers to which he hopes will lead him to solve the main one. A favored form of the basic question about literature is "What is the meaning of this play, this novel, this lyric poem?" This question branches out into other questions of meaning: of words, phrases, themes, plots, symbols, stylistic devices . . . . Questions of meaning are, as the linguistic philosophers have shown and as everyone now knows, very difficult to answer. In part this is because the verification of problems of meaning is not empirically possible, as meaning is an abstraction at a certain remove from the events under examination. Disagreements about meaning, about the interpretation of literary works, abound in literary study. In a sense every interpretation is correct or at least justified since it may be supported by a proper selection of evidence and since there is no established priority governing the evaluation of evidence. Thus literary interpretation generally depends for its effect on the persuasiveness with which the selection of evidence is presented and it usually relies for its acceptance on a certain set of beliefs or expectations common to the interpreter and his audience. For example, the question "What does *Hamlet* mean?" has been answered in this century by reference to the possible incestuous relationship between Hamlet and Gertrude and Hamlet's supposed Oedipus complex, which in turn have been traced to some emotional difficulties in the playwright. But other meanings of Hamlet have been successfully defended which are supported by a different selection of evidence.

As is well known, there can be no progress in interpretation. Explications survive for as long as the willingness to believe the theory on which they are founded persists. When the winds of critical doctrine change, what was previously acceptable—theory, interpretation, evidence—is swept away to be replaced by the newer thing. This is a discouraging state of affairs but one to which literary scholars have become adjusted. Their way of adjusting to this fluid and unstable situation is by the reduction of big problems to little ones, by the conversion of *why* questions to *how* and *what* questions.

Preliminary to any inquiry about the meaning of a given work is usually a set of subordinate questions, some of which may seem rather remote from the main event. Thus, the study of *Hamlet* implies the study of the medieval theatre, beliefs about lunacy, ghosts and family relationships, the sources of this particular play, the shape and appointments of the Elizabethan playhouse, Shakespeare's life, his philosophy as it is reflected in the speeches of his characters and in his imagery, his language as it differs from or corresponds to the language of the playwrights and writers of his time, and an innumerable list of sub-questions. Presumably, when all the evidence is in on these lower-echelon matters, the main question—"What is the meaning of *Hamlet*?"—can be tackled, unless someone comes along with a critical theory that denies the possibility that plays or other literary artifacts can have meaning, apart from mere existence. The words of a modern poet record this position of critical nihilism: "A poem should not mean but be."

Without an unceasing concern for the ultimate necessities, the questions of meaning and value, any study is in danger of becoming merely the trivial sorting of artifacts, the solving of puzzles or riddles no more significant than a newspaper crossword. In other words, literary scholarship, computerized or traditional, must be informed by this concern for what literature is and means and for the things that literature springs from and tries to illuminate. Computer scholars are more vulnerable to such a danger than traditional scholars because the traditional tools of research—cards, files, pencils and typewriters—do not exercise the dangerous and autonomous fascination that the electronic data-processors do. The computer study of literature always threatens to take over the scholar, who becomes seduced by the ease with which it can do certain things into abandoning his real goals and responsibilities. At the same time, he is subject to another sort of accusation which is quite the opposite. If he uses a computer, he is expected to solve all the outstanding problems of literature, simply because the real accomplishments of the computer and the efforts of the manufacturers' public relations men have accustomed the public to expect decisions, solutions and miracles from the computer, as a matter of routine. Anything less counts as a failure. These and other

jeopardies face the literary scholar who has turned for help to a computer.

The problem is truly paradoxical. The traditional literary scholar cannot solve the great problems unless he first solves the small ones. These invariably consist of the accumulation and compilation of data, minute in size and immense in quantity. If he immerses himself in these, he is very likely to lose sight of his original purpose. If he does not, his conclusions are mere baseless speculations. To this paradox, the computer can bring a solution because of its ability to undertake the drudgery required for answering the subordinate questions.

Nonetheless, it is difficult to escape the conclusion that computer-assisted literary scholarship has until now been woefully conservative. It has done little to exploit the machine's genuine possibilities for qualitative innovation. It has largely limited itself to the mere quantitative aspect. These efforts will doubtless earn some praise as the results become available and useful to the community of scholars, but they will not inspire other scholars to emulation because the results are not truly inspiring or exciting. Not until the computer scholar turns out results which diverge sharply from what has been done before will he earn the respect and interest of his traditional colleagues.

One explanation of the conservative nature of the computer-assisted projects has to do with the relationship between the scholar and the machine. He has learned to think of it as a highly efficient but brainless clerk—despite everything written about artificial intelligence. Therefore he has entrusted it with merely clerical tasks. Moreover, he has usually employed an intermediary to convey his instructions because he is not himself sufficiently conversant with its language to do so himself. In a sense, therefore, he is doubly dependent and doubly limited: he has a foreshortened view of the computer's abilities and he must depend on the understanding of another person to express his needs. He must free himself of both these limitations if he wishes to make his scholarship creative. Both, it seems to me will yield to the one cure: the scholar must learn to be his own programmer. That is axiomatic. He cannot learn what the computer can do if he has to ask another to interpret for him. With the development of new high-level languages like SNOBOL, competence in which can be acquired even by the stiff reflexes of the middle-aged scholar, though not without effort, there can be no excuse for remaining technologically illiterate. The scholar who familiarizes himself with the means of communicating with his computer will learn at the same time how extensive are its possibilities, how untried its opportunities.

The aims of humanistic scholarship, according to the recent words of a well-known classicist, should be primarily educational. They should, that is, instruct the scholar and enlighten his instruction, especially in the sense that a knowledge of the past can help one to judge the present. If the fulfillment of the literary humanist lies in this sort of activity, the computer properly used can make a considerable contribution.