ARRAS and Literary Criticism

John B. Smith

Department of Computer Science
The University of North Carolina
Chapel Hill, North Carolina 27514
919-962-5021

Abstract

This paper has two major objectives: to describe ARRAS -- an existing computer system for conversational, open-ended studies of large texts -- and to consider the implications of using such a system for literary studies. ARRAS permits the scholar to display, immediately, the vocabulary of a text, the contexts of words, the spatial distributions of words or groups of words, and the locations of logical contextual patterns. Used within the CMS environment, the scholar may combine the information produced by ARRAS with his or her own comments in a word-processing context as well as communicate with other scholars who have access to a computer network. The result is a new interpretive environment.

The implications of this new environment are several: it will affect the scholarly questions that are asked, the answers that are accepted, as well as the experience of interpretation. Future development will require not only other systems such as ARRAS and large archives of texts, but also a new critical theory that can link these tools, materials, and conceptualizations with the mainstream of critical theory.

ARRAS

This paper has two major objectives: to describe ARRAS -- an existing computer system for conversational, open-ended studies of large texts -- and to consider the implications of using such a system for literary studies.

ARRAS (ARchive Retrieval and Analysis System) is an interactive system designed to give fast, flexible access to texts, such as novels or plays, selected from a data base of texts. It is intended to be used by the scholar via terminal from his or her study, office, or usual workplace. The scholar selects the text to be considered, refers to the printed copy when appropriate, but refers to the terminal for concordance information, lexical information, graphical distributions, or other forms of analytic data. When used within the IBM CMS (Conversational Monitor System) environment, the scholar may combine the information produced by ARRAS with his or her own writing in a word-processing context and, in some cases, exchange drafts with colleagues at other institutions via computer network. Together, ARRAS, a collection of texts, and CMS constitute a new working environment for the academic.

ARRAS consists of three major components: a data base of texts, an integral self-instruction system, and the analytic system itself. Texts, either typed into the computer or scanned optically, are presented to ARRAS virtually as they appear on the printed page except for a few insertions that mark divisions. They are then transformed into the required ARRAS data structure and stored in the CMS file system for later use by ARRAS.

The scholar normally initiates an ARRAS session by signing onto the CMS system and typing the word, *ARRAS.* After presenting a logo, ARRAS awaits the scholars command. If the scholar does anything other than type a correct command, ARRAS offers help. The help system includes an introduction that provides an overview of the system, an explanation of the general form of the command language, and descriptions of each command word. Thus, the scholar may use ARRAS without having to rely on printed documentation.

The heart of the system is the analytic system. Time does not permit me to describe all of ARRAS' functions or to describe all possible uses of them, but I will try to provide a sense of the system's capabilities by describing a half-dozen of its more frequently used commands. Let me emphasize, however, that ARRAS is intended to be open-ended: the basic tools it offers can be used for many different kinds of studies and the system is designed so that new capabilities can be added without compromising its integrity. I will illustrate ARRAS by describing how it might be used for a thematic study of a novel; the foils I will show apply to James Joyce's *A Portrait of the Artist as a Young Man.*

One might typically begin such a study by considering the author's vocabulary. ARRAS can display the entire vocabulary, in alphabetic order and with the frequency of occurrence for each word. More often, the scholar will ask for some portion of the alphabetic sequence, such as all the words in the alphabetic range *fire - firez.*

After examining the author's lexicon, the scholar may ask for contextual information for words he or she is interested in by asking ARRAS to display a concordance for a given word. Such a request is answered by ARRAS immediately, usually in less than a second. The context normally supplied is each full sentence in which a word appears, but the scholar may ask for more or less context, measured in units of *words, sentences, paragraphs, chapters, volumes, pages.* Since ARRAS can also supply context for a single occurrence of a word, the scholar may wish to see only a very brief concordance (perhaps three or four words on each side) and then examine in more detail those places in the text where his or her interest is piqued.

Many studies, however, will wish to focus not on individual words but on associated groups of words. ARRAS provides very powerful categorization support. One simply informs ARRAS that a group of words is to be considered a category, supplies a name for the cluster, and ARRAS will keep-up with the set. In any command where a text-word may appear, a category name may be used. Thus, a concordance for a category produces the contexts for all words included in the category. However, categories may consist of other category names. In ways that will become

clearer later, when I describe the *configuration* command, a major part of many analyses using ARRAS will be the development of an evolving hierarchy of categories. To support such studies, ARRAS offers a number of commands for saving, retrieving, and modifying the collection of categories.

To help the scholar gain a spatial sense of a word or category as it is used by an author over the course of a long text, ARRAS offers pictorial distributions of words or categories. A request for a distribution results in a bar graph in which the text is divided into fifty segments of equal length (for the purposes of display; ARRAS could produce a graph of any desired resolution). Each column of the bar-graph indicates the number of occurrences of the word or the set of words for that text interval.

After examining the contexts for words or categories, after looking at their distributions over the text, the scholar may next wish to search for locations where some specific combination of words or categories of words occur. ARRAS permits the scholar to search for contextual patterns using full Boolian logic. For example, he or she may ask ARRAS to find all the places where any word in the *fire* category occurs in the same sentence with any word in the *water* category but where there is no word from the *religion* category in the same paragraph. The result of the search is a set of locations; this set may be named and becomes a category analogous to a category of words.

Categories of locations can be used like any other category -- in a *concordance* command, a *distribution*, or in another *configuration* search. Thus, one may look for patterns defined in terms of words, but then move on to look for patterns of patterns, patterns of patterns of patterns, ... For example, one could look for syntactic patterns in which the categories represent parts of speech. The resulting categories would represent sentences of a particular form. A subsequent search in which the search expression included categories produced by the first search would locate paragraphs composed of logical patterns of sentence-types.

After viewing the authors vocabulary, exploring his or her use of key concepts, viewing distributions of words or categories, searching for logoical configurations, etc., the scholar may wish to pause, make notes, add several paragraphs to an article being written, or save a copy of some of the information provided by ARRAS for future use. To do these things, he or she instructs ARRAS to save a copy of whatever is displayed on the terminal screen in a CMS file. The scholar may then leave ARRAS and work within the CMS environment. There, he or she can edit the file of notes or journal article, add a paragraph describing the insight, add a graphic distribution or concordance listing to document the point, etc. Should the scholar's computer system be linked to BITNET, he or she could also exchange drafts of the article with a colleague at another institution as easily as he or she can do so with a colleague using the scholar's own computer system. When the CMS activity is completed, the scholar may return to ARRAS and pick-up the analysis here it was left off -- with all analytic categories intact.

Implications

While this brief summary omits a number of ARRAS functions, I hope it will provide a feel for the system. Let me now step back and consder the implications of a system such as ARRAS for literary studies.

A setting in which the scholar may refer to conventional published materials, may turn to the computer for immediate analytic information, may do word-processing, and may communicate quickly and easily with a nation-wide community of scholars is a new kind of intellectual environment. It is not just a collection of tools but a critical mass of components that, symbiotically, will have greater consequence than the sum of their individual contributions.

The technology is unlikely to remain neutral. The most conventional component of ARRAS is probably its concordance feature. Concordances are familiar research tools. Scholars are comfortable with them and they know what to use them for. ARRAS supplies conventional concordance information, but it does much more. Simply by providing immediate control over context, it makes the scholar aware that he or she is working with a tool that is different from the

traditional published form. As a second instance, the categorization feature, when used with an external thesaurus or set of categories, can accomodate differnt principles of lemmatization. Thus, many of the decisions that have been debated by scholars regarding concordance structure are either circumnavigated or made variable by the interactive nature of the computer.

More important, though, the computer is likely to influence the scholar's thinking. That is, its abilities to supply information immediately and to search for patterns too subtle and too diffuse to be perceived directly are sure to influence the questions asked, the anwers accepted, as well as the crtitical experience, itself. With the intuitive notion of what a powerful interpretive system looks like, let us now turn to several issues that relate more directly to critical theory.

ARRAS and other simililar systems will permit scholars to approach questions that until now have been impractical or would have required many years. Let me illustrate with several examples. First, scholars will be able to ask questions that deal with close, intricate textual patterns. I am reminded of Caroline Spurgeon's study of Shakespeare's imagery. By cataloging and categorizing each image in Shakespeare, Spurgeon was able to show that the Bard used imagery to establish a pervasive atmosphere or verbal backdrop for many of his plays. The audience may have been only subliminally aware of these patterns, but Spurgeon argues convincingly that these patterns have much to do with a play's overall affect. In the future, studies of this sort will be greatly facilitated by the computer.

A second set of questions concerns longitudinal trends that can be traced only by considering in close detail a large number of works covering some period of literary history. When associated with an archive of texts, ARRAS will make such studies much more feasible. As one example, the University of Chicago will soon offer ARRAS and a data base of French belle lettres to scholars in French studies. As typical examples, they will be able to consider, precisely, how the concept of *liberty* or how the French familiar form of address changed during the period of the French Revolution

As the questions scholars ask change, so will the answers they accept. That is, the computer will change the notion of what constitutes adequate demonstration of a point. Today, scholarly arguement in literary studies rests of three epistemic pillars: logic, citation of authority, and citation of examples. No doubt logic will remain the basis of argument, and we shall probably continue to cite the opinions of those who agree with us. But the computer will make us see that citation of a textual instance that illustrates a generalization is inadequate as proof. Since the computer can show us each and every instance of a word or pattern in a text, we may now include the notions of pervasiveness and adequacy in our concept of argument. That is, we can not only note the existence of some textual pattern, but we we may also ask to see how frequently that pattern apapears.

Should the pattern appear only once or twice, that may be significant; but it is a point different from evidence of the generality of some assertion.

The computer will also help us see what is left out of an interpretation as well as what is included. For example, if someone is doing an imagery study such as the Spurgeon study, he or she may view the author's entire vocabulary. In doing so, the interpreter may look to see that the set of images used in the study is complete. That is, the scholar can see, precisely, what words are not included in the image set as well as those that are imcluded and, thus, insure that substantive words that are omitted are done so intentionally.

A third way in which ARRAS may influence the scholar's thinking is in the very nature of the interpretive experience, itself. Wolfgang Iser has made us aware of the phenomenology of reading. He points out that the experience is different the second time we read a text from the first. *Surprise*, for example, is replaced by a heightened awareness of the unfolding of pattern. While Iser does not develop the point, we can extend his distinction to include a third kind of experience: the detailed study of a literary work. Multiple readings, patterns consciously sought and observed, notes taken or diagrams drawn to enhance recollection and perception, all lead to an experience that is different from both first and second readings. I call the critics experience the *phenomenology of interpretation.*

Let's go one step further -- the phenomenological experience of the interpreter using ARRAS will be different from that of the conventional reader who relies on memory and the printed text, alone. That is, the interpreter who uses the computer to "look down" on an entire work or group of works can recall immediately and exactly all instances of a word, phrase or pattern. He or she can use the computer to raise to consciousness patterns too subtle or too diffuse to be perceived directly. This interpreter who uses the computer to augment memory and perception will have an aesthetic experience that is different in kind from that of the conventional interpreter. The situation is analogous to microphotography or time-lapse photography: these technologies permit us to see the world around us in ways that are qualitatively different from everyday experience. The computer will permit us to experience literature in ways not possible without its help.

ARRAS, of course, is just one step on the path to this future of possibilities. What do we need to realize fully what is only hinted at here? In the few minutes left, let me sketch, briefly, the three major tasks I see for those interested in bringing the computer into the mainstream of literary scholarship. If the computer is ever going to be used extensively, we need three large classes of resources: archives of texts in machine-readable form; powerful, easy to use computer systems; and a critical context.

Scholars must be able to get copies of the texts they are interested in exploring in a form that can be read by the computer. While the optical-character reader will reduce substantially the costs of encoding texts (probably on the order of 60%), the cost is still not inconsequential and many institutions cannot afford the $90,000 to purchase one of these devices. Institutions, such as Oxford University, that have begun to encode texts are doing so in a hap-hazard way, with no assurances of accuracy or format. The countries whose scholars study English-language texts need a coordinated commitment to encode with a specified level of accuracy and specified formats the major literary corpora. Such a commitment will no doubt require several ten's of millions of dollars, but as a world resource, that is not an extravagant sum.

Second, scholars need powerful, flexible systems. I believe ARRAS is a start in this direction, in its flexibility, its general power, in its interactive design, and in its internal self-instruction system for external or network access. We need other such systems that perform other tasks. While I would not endorse any imposed restrictions on design, informed cooperation should be sought to permit these systems to work in concert.

Finally, we need a theoretical basis for the studies that will result from the use of these resources. I have tried to give some indication of a few of the ways the computer will influence the thinking of the scholar who uses it. I have only scratched the surface here. Whether we intend it to or not, the computer will lead to new forms of scholarship and interpretation. For lack of a better word, I have termed such studies *Computer Criticism.* Without foundation studies, the results will be *ad hoc* and arbitrary. We need to describe in detail the conceptual point of view of *Computer Criticism* and to chart the relation that point of view has with the major currents of twentieth century critical theory. By doing so, we can profide a coherent basis for sustained development. Until this is done, computer studies of literature will remain peripheral.