Signs, Symbols and Discourses: A New Direction for Computer-Aided Literature Studies

Mark Olsen

ARTFL Project, University of Chicago, Chicago, IL 60637, USA
e-mail: mark@gide.uchicago.edu

Abstract: Computer-aided literature studies have failed to have a significant impact on the field as a whole. This failure is traced to a concentration on how a text achieves its literary effect by the examination of subtle semantic or grammatical structures in single texts or the works of individual authors. Computer systems have proven to be very poorly suited to such refined analysis of complex language. Adopting such traditional objects of study has tended to discourage researchers from using the tool to ask questions to which it is better adapted, the examination of large amounts of simple linguistic features. Theoreticians such as Barthes, Foucault and Halliday show the importance of determining the linguistic and semantic characteristics of the language used by the author and her/his audience. Current technology, and databases like the TLG or ARTFL, facilitate such wide-spectrum analyses. Computer-aided methods are thus capable of opening up new areas of study, which can potentially transform the way in which literature is studied.

Key Words: computer-aided literature studies, literature, literary theory, structuralism

Computer processing of textual data in literary and historical research has expanded considerably since the 1960s. In spite of the growth of such applications, however, it would seem that computerized textual research has not had a significant influence on research in humanistic disciplines and that literature research has not been subject to the same shift in perspective that accompanied computer-assisted research in more social science oriented disciplines, such as history. This is vital since the majority of current practices have failed to capture the support and imagination of colleagues who do not use computers. We must address the issues surrounding the general failure of our discipline to have a significant impact on the research community as a whole. An important corrective is that computer methods in textual analysis allow scholars to ask new questions which do not correspond to the traditional notions of reading texts. Thus, I suggest that a shift in the theoretical orientations of computer-assisted textual analysis may lead to a more prominent role in the mainstream of literature.

In spite of the investment of significant amounts of money and time in many projects, the role of electronic text in literary research remains surprisingly limited. For many years, the typical defense of the potential of computer analysis of text was argued from the lack of resources. If only we had X – where X could be more texts, better software, more processing power, more disk storage, and so on – then the utility of these methods would be demonstrated. This is no longer the case for many disciplines. Large bodies of textual data, such as the Thesaurus Linguae Graecae (TLG) on CD-ROM or the ARTFL database, combined with powerful workstations that can provide sophisticated access to these databases at low cost suggest that the problem of impact is not necessarily resource oriented. It is equally true that the failure of computerized textual analysis to influence researchers can no longer, now that the wordprocessor and bibliographic database are standard equipment for literary scholars, simply be blamed on "computer phobia."

Recently, some researchers working with elec-

Mark Olsen received a Ph.D. in French Revolutionary history from the University of Ottawa and is Assistant Director of the ARTFL Project.

Computers and the Humanities **27**: 309–314, 1993.

tronic texts have wondered openly about the continued failure of computerized textual analysis to provide results that are of general interest to researchers in a field. Dee Clayman, a classicist at CUNY, recently argued in a paper provocatively titled "The Quantitative Analysis of Text: Why has it Failed?" that

> it [quantitative analysis of text] has not yet produced results that other scholars consider to be reliable or useful. That is not to say that our work so far has been in error, or in vain, but rather that we have not yet succeeded in convincing very many other people to believe in it.[1]

Indeed, she goes on to argue, graduate students who engage in computer projects quickly learn that it is a form of "professional suicide," rather than something that can benefit either their research or careers. Thomas Corns puts the matter more bluntly, suggesting that "there is no substantial body of achievement in the field of computer-based literary criticism in English studies." A scan of the most prestigious journals of criticism or history reveals that few scholars use electronic text at any level in their work. Corns notes, for example, that no work using computers in literary analysis appeared in two important mainstream periodicals, *Yearbook of English Studies* and *Review of English Studies* in the 1980s.[2] Similar conclusions can be drawn for other areas and disciplines which use textual information as a major element of their research.

We have seen the rise of specialized journals – including *Computers and the Humanities* (*CHum*) and *Literary and Linguistic Computing* (*L&LC*) – where the results of such work appear. There is little cross-over from these journals to mainstream publications. Indeed, the degree to which computer-assisted textual analysis has not had a significant impact was underlined in a recent number of *New Literary History* devoted to "Technology, Models, and Literary Study" which does not contain a single citation to works published in journals like *CHum* or *L&LC* or other studies using computers to perform significant text processing. This systematic omission does not seem to be the result of ignorance. Richard Lanham cites the existence of the TLG and Packard Humanities Institute CD-ROMs in his essay on "The Electronic Word: Literary Study and the Digital Revolution," while William Paulson makes reference to the ARTFL database in his "Computers, Minds, and Texts: Preliminary Reflections."[3] It seems clear that the current generation of computer-assisted literary analysis has not captured the attention of scholars who, as in the case of authors examining computation and literature such as Lanham and Paulson, are very well disposed to the technology and see an important future role in literary production, distribution, and analysis of computers.

The reaction of well-known and influential scholars to the use of computers in literature research has been frequently negative. Jeremy Popkin, for example, in a recent discussion of trends in the intellectual history of the French Revolution, justifiably criticizes much of the work of the Centre de Lexicologie politique de Saint Cloud as found in the journal MOTS: Mots, Ordinateurs, Textes, Société, suggesting that it has been

> too narrowly focused on a single text or a single author to provide much evidence about changes in usage over time, and the results have been expressed in a complex jargon that makes them unintelligible to most working historians.[4]

Rosanne Potter argues that critics using computer technology have tended to write to more technically oriented audiences rather than to scholars in their own disciplines:

> Many readers of criticism do not know what can be done with computing because, until now, most literary users of computer technology have not written the essays they could have written about their work. Beautifully clear essays about literary critical data have gone unwritten because of the necessity of using a generally accepted scientific style to meet the standards of reviewers at journals interested in computing research. As a result, many computer critics have written themselves out of the range of their natural audiences.[5]

Potter's program in her volume is to publish essays that do not have the dense statistical and linguistic terminology that she believes is the most formidable obstacle to general acceptance of computing in literary criticism. While better identification of the audience of textual research is important – and the essays in Potter's volume make a clear step in that direction – it seems overly simplistic to suggest that the problem is a matter of clear writing.

There seem to be several reasons for the consistent failure of computer methods to engage the research community in text-oriented work. The

failings of computerized text analysis are not based on technical computer issues – in terms of software or hardware limitations – but on theoretical and methodological issues. Practitioners of computing in the humanities have frequently developed profound technical expertise, but this knowledge is not translated into interpretive strategies. This is due to an apparent lack of a corresponding shift in theoretical models and methods to match the new technology. Rosanne Potter identifies this problem in her own volume, expressing surprise at the limited nature of all of the submissions to her volume:

> I had expected [she writes] some broadly descriptive essays on the qualities of different genres or periods, but at the moment most critics who are doing serious literary computing [whatever that means] are contrasting one or two writers, comparing two or three works by the same author, or concentrating on one work at a time.[6]

She then goes on to mention Burrow's recent study of Jane Austin's novels as an exemplar of the narrow focus of most literary computing research.[7] The same tendency is found in the "Special Section on Literary Criticism and Computing section of a 1990 number of *Literary Linguistic Computing*, which includes studies of three tagged novels by Eliot, Shakespeare's style, half a dozen poems, and the first fragment of the *Canterbury Tales*.[8] Both of these recent collections make important statements of theoretical intent, which attempt to justify the selection of methods and datasets.

The dominant theoretical models underlying much of this research are based on stylistics and a modified kind of reader response. These models assume the existence of structures which are not immediately apparent to the reader and which are vital elements of the literary effect of a text. The most frequent method typically involves plotting the distribution of textual elements – ranging from simple words to any variety of features tagged by a human or a machine – across a text. Teresa Snelgrove asserts that quantitative analysis of tags she assigns to a text allows her to

> perceive that elusive symmetry of a text which is observable only when we attain that vantage point by which we can see the object as a whole.[9]

Paul Fortier argues that computational stylistics can be a means to isolate the ways in which an author modifies the perceptions of the reader by examining the element of surprise found in incongruous lexical constructions.[10]

The relationship of stylistics and reader response theory is problematic, but the principal focus of both approaches is to measure the degree of literary art discertible in a text or achieved by an author. The computer can, of course, be used as a rapid means to produce a concordance or to allow the scholar to track down words associated with particular topoi or images, but it would seem that most proponents of computer-assisted criticism aspire to something more important: the systematic discovery of how a text achieves its literary effect. These structures escape the attention of a careful reader, and can thus be uncovered only by systematic means. A second benefit, frequently mentioned by scholars using computers, is the high degree of verifiability of computerized and/or quantitative criticism.

It is the concern with the achievement of the literary effect, measured in the text itself (if that is possible) or in its presumed impact on a reader (hypothetical or real), which seems to be the single greatest problem. Failure to develop alternative perspectives has led to consistent frustration with the overwhelming complexity of language and text. The difficulty in developing minimally effective morphological analyzers and syntactic parsers has led to a proliferation of inelegant "work arounds" or forced researchers to code only small amounts of data. As Potter suggests:

> most computer-assisted critics do not expect artificial intelligence to find most of the syntactic features easily recognized using human intelligence. These critics just type the text into the computer's memory along with specific markers for each complex syntactic feature.[11]

The most successful computer methods in literary and language based research have treated the simplest, most readily identifiable elements of language, with no significant element of analysis performed by the computer. Concordance generation has been, since the 1960s, the paradigm of computer applications in the humanities. Authorship studies have relied on measures of very simple textual features, like word lengths or type-token ratios. Word frequencies and phonetic patterns, structures which can be easily identified by computer programs, tend to dominate computer-aided criticism, even though there is general

dissatisfaction with the limited nature of the results of this kind of research. Attempts to move beyond such simple applications have met with more limited success, because the complexities of text and language have eluded simple measures based on word frequencies or content analysis. Van Peer sets out this "paradox" by arguing that

> the illocutionary force and meaning aspects of a text are much harder to quantify than issues of sound and grammar [. . . but it is] exactly these more abstract aspects of linguistic organization, such as meaning or force which are at the core of textuality and indeed of literature.[12]

Some critics of computer-assisted analysis of textual data suggest that, after all the programs are finished running, computer analysis of texts does not tell us much we did not know before, and worse, may not prove the important matters any better than a close reading.

By attempting to ask fairly traditional questions of traditional texts, computer applications in literary and language based research have failed to move from a curiosity to an important and respected position in these disciplines. By contrast, quantitative social, political and economic history used computer technology to ask new questions and to develop new methods. Indeed, the computer fit nicely into a shift away from political and event based history, to the history of social phenomena and the long term, *la longue durée*. The traditional object of historical research, the serial history of events, has not been nearly as revolutionized by the new methods, which are frequently based on computer technologies. It seems that computing in textual research has not seen a corresponding shift in perspective or the development of new objects of research.

To this historian, the most surprising element of this problem is the relatively limited acceptance of critical models that stress intertextuality and sign theory among those interested in computer analysis of text. This is not to say that the structuralist and post-structuralist critical theories are not talked about in relationship to computer-assisted textual research, but that few systematic attempts have been made in this area, combining software, data selection, and theory in interesting ways. Such theories have, to my mind, several attractions to computer applications. The first is that they encourage research design that exploits the strongest points of computer technology, the high speed access and analysis of large amounts of data. The individual text and author, in these models, become less important than the manipulation of signs. The second is that concentration on levels of intertextuality, through language and signs, avoids the more complex elements of textuality which have proven to be far more elusive in computer-assisted research than had been hoped.

Roland Barthes points to the importance of viewing literary texts in their linguistic or semiotic context. Literary texts cannot be delimited simply by their covers, since they are composed of codes and discourses drawn from many other texts. He writes:

> We know now that a text is not a line of words releasing a single "theological" meaning (the "message" of the Author-God) but a multi-dimensional space in which a variety of writings, none of them original, blend and clash. The text is a tissue of quotations drawn from the innumerable centers of culture. [. . .] Did he [the Author] wish to express himself, he ought to at least know that the inner "thing" he thinks to "translate" is itself only a ready-formed dictionary, its words only explainable through other words, and so on indefinitely.[13]

Barthes is arguing for a critical evaluation that runs considerably more profound than simply tracing "influences," since he suggests that it is the language of the text that is central. Barthes' notion of language is free-floating, for he does not see that signs can ever be given a fixed meaning. Signs, however, may be often fixed by their relationship of social power relations. While it is clear that the author does not control or invent the meaning of signs at a particular time, Barthes' radical solipsism ignores the social construction of meaning. Foucault grounds knowledge and language in the deployment of power through the social structure.[14] The identification of social semiotics as an important element of the constitution of language through expression of social power relations is not limited to post-structuralist critics. Sociolinguists, such as M.A.K. Halliday, have argued that

> language plays a central role, both as determiner and determined: language is controlled by the social structure, and the social structure is maintained and transmitted through language.[15]

Feminist critics of literature and language have suggested that the unconscious structuring of meaning is the primary means by which gender power relations are linguistically encoded and perpetuated. Ann Rosalind Jones writes,

> Symbolic discourse (language, in various contexts) is another means through which man objectifies the world, reduces it to his terms, speaks in place of everything and everyone else – including women.[16]

Barthes' declaration of the importance of linguistics in literary criticism and his reduction of the intentions of the author in understanding a text points to an alternative model of literary computing: computers can be used to isolate and examine the symbolic universe in which an author writes.

The computer is well suited to examining words, expressions, and other clearly defined segments of texts. Indeed, it might be argued that the computer is an ideal semiotic machine, for it can examine vast numbers of signs rapidly, comparing sign use in defined blocks of text. Finally, systematic analysis of sign or language use across large blocks of text is amenable to both qualitative and quantitative analysis. The image of the Fortunate Islands (later the Canaries) or "women's anger" can be examined systematically in a large database without resorting to quantification at all, simply by reading and comparing descriptions across many texts over many years. By contrast, one can examine the changing meaning of words that constitute the discourse of gender over long periods of time using quantitative methods that are well known and understood. Systematic examination of very common words, such as *femme* and *homme* reveal startling shifts in the linguistic encoding of gender relations of which contemporary users of the French language would have been unaware. Preliminary results from my current work on the ARTFL database suggest that women were not categorized by age in the seventeenth century while age became a primary modifier of terms denoting woman by the end of the nineteenth century.[17] It is clear that computerized textual analysis, based on available technology and software, can support new forms of literary analysis.

It would seem that the use of computers to analyze the linguistic and symbolic environment – the collective and social elements of language – in order to understand individual texts and rhetorical stances, suggests that computer analysis of text should play a central and well defined role in our understanding of text. This role fits the strengths of computational analysis and the theoretical models used to inform research on text and language. This requires a move away from the traditional literary emphasis on the author's intention in writing the text and the stylistic construction of individual texts. Modern critical theory, particularly of the structuralist and post-structuralist traditions, provide precisely that theoretical opening.

This model of the role of computer-assisted text analysis may parallel the "history of meanings." Rather than the study of single ideas, thinkers, or problems, intellectual and cultural history using electronic text can be thought of as the systematic investigation of the development and transformations of "meaning systems." Meaning here is taken in its structuralist sense, as culturally defined linguistic and non-linguistic symbols and codes that are used to comprehend and organize the external world. By stressing the "constructedness" of human meaning systems, the historian can examine the assumptions, limitations, changes, and contours built into shared systems of signification. There are, it would seem, two distinct levels of analysis in a history of meaning: the examination of limited numbers of texts/speech acts and, more broadly, the attempt to establish patterns of common usage, which may be related to *mentalité*. The difference is one mainly of perspective. In the first case, one would compare the discourses of defined social or political groups. In the second, one would attempt to map the language common to very large populations. The literary critic may select different characteristics by which to define groups, such as examining the discourses of male and female writers or between genres of a particular period. The perspectives merge when assessing changes in a particular discourse, which may require knowledge of underlying structures.

Authors function within symbolic universes of which they can only be partially conscious. If I am correct in assuming that computer analysis of large bodies of material for relatively simple constructions reveals elements of discourse that are beyond the control or awareness of an author, then studies of individual authors and texts will be rendered more intelligible by highlighting the way in which an author uses, modifies, or rejects contemporary symbols. Indeed, selection of interesting or important rhetorical shifts or manipulations would seem to presuppose careful identification of general patterns of discourse. Knowing what

language is available to an author – the symbolic universe in which he/she operates – is a vital part of interpreting the intentions and limitations of the text.[18] This might be accomplished by comparison of statistical evidence drawn from two corpora, such as the works of a particular author and works published during the period in which he or she was active, or simply bringing to the reading of a text detailed and verifiable knowledge concerning contemporary language and representation of selected themes without further quantitative analysis. The relationship between two radically different modes of examining text will not be solved in a single formula because the role of the author and the nature of individual texts in creating meaning are contested. The literary critic cannot write about the particular instance of writing without understanding the collective background.

None of this is to suggest that computer-assisted analysis of single texts or individual authors is either pointless or illegitimate. Indeed, scholars like Clayman and Van Peer remain convinced that stylistics and other forms of computer analysis of text data result in useful and interesting findings. I suspect, however, that computers will make a much larger impact on literature when new questions and data types are addressed and when computers are used to do things that humans are either incapable of, or that are simply too time-consuming. Implicit in this suggestion is the notion that computation in itself does not constitute a workable model of literary computing. The traditional notions of textual analysis are not well suited to computer development. Thus, a reorientation of theoretical models underlying computer-assisted textual research will exploit the strengths of current computer technology and provide an important corrective to the traditional concept of reading texts. Such a change of optic will ensure a central role for computer analysis in many areas of textual research.

Notes

[1] Dee L. Clayman, "The Quantitative Analysis of Text: Why has It Failed?" (unpublished paper presented to CREDO, 1989), p. 1. See also Willie van Peer, "Quantitative Studies of Literature: A Critique and an Outlook," *Computers and the Humanities*, 23, 4/5 (1989), 301–308.

[2] Thomas Corns, "Computers in the Humanities: Methods and Applications in the Study of English Literature," *Literary and Linguistic Computing*, 6 (1991), 128.

[3] In *New Literary History*, 20 (1989), 265–303.

[4] "Recent West German Work on the French Revolution," *Journal of Modern History*, 59 (1987), 748.

[5] Rosanne Potter, ed., *Literary Computing and Literary Criticism: Theoretical and Practical Essays on Theme and Rhetoric* (Philadelphia: 1989), p. xviii.

[6] Ibid., p. xix.

[7] Corns (p. 127) points out that Burrow's study, interesting in its own right, highlights the problems surrounding computer criticism: "its reception and its influence in humanities computing circles reflects as much as anything the solitariness of its achievement: a wholly statistical account published by a major university press."

[8] *Literary and Linguistic Computing*, 5 (1990), 221–47.

[9] Teresa Snelgrove, "A Method for the Analysis of the Structure of Narrative Texts," *Literary and Linguistic Computing*, 5 (1990), 221.

[10] Paul Fortier, "Analysis of Twentieth-Century French Prose Fiction: Theoretical Context, Results, Perspective," in Potter, ed., *Literary Computing*, p. 92.

[11] Potter, p. xxv.

[12] Van Peer, p. 306.

[13] Roland Barthes, "The Death of the Author," in Stephen Heath, ed. and trans., *Image, Music, Text: Roland Barthes* (New York, 1977), p. 146.

[14] After May 1968, Foucault moved from his concern with archeologies – unchanging structures of knowledge which can "explode" suddenly – to the problem posed by "power/knowledge." See the essays and interviews collected and translated in *Power/Knowledge* (New York, 1980).

[15] M.A.K. Halliday, *Language as Social Semiotic: The Social Interpretation of Language and Meaning* (London, 1978), p. 89. See also M.A.K. Halliday and Ruqaiya Hasan, *Language, Context, and Text: Aspects of Language in a Social-semiotic Perspective* (Oxford, 1985).

[16] Ann Rosalind Jones, "Writing the Body: Toward an Understanding of *l'écriture féminine*," in Judith Newton and Deborah Rosenfeld, eds., *Feminist Criticism and Social Change: Sex, Class and Race in Literature and Culture* (New York, 1985), p. 87.

[17] "Gender Representation and *Histoire des Mentalités*: Language and Power in the Trésor de la Langue française," in *Histoire et Mesure* VI (1991): 349–73. Similar work can be performed to examine the evolution of political discourse. See, for example, Mark Olsen, "Enlightened Nationalism in the Early Revolution: The *Nation* in the Language of the *Société de 1789*," forthcoming in *Canadian Journal of History* XXIX (1994) and "The History of Meaning: Computational and Quantitative Methods in Intellectual History," in Daniel Woolf, ed., *Intellectual History: New Perspectives*, volume 6 of *Journal of History and Politics* (Lewiston, N.Y.: Edwin Mellon Press, 1989), pp. 121–54.

[18] See Mark Olsen and L.-G. Harvey, "Contested Methods: A Discussion of Daniel T. Rodgers" Contested Truths, Keywords in American Politics Since Independence," *Journal of the History of Ideas*, 44 (1988), 653–68.