Electronic Texts: A Promise for Humanities Research

by Kurt De Belder

Originally appeared in: *Academic Computing and Networking at NYU*, Vol. 3/No.4, May 1993.

Kurt De Belder, Western European Literatures and Languages Librarian at Bobst Library, is a member of the Library's Task Force on Electronic Texts

Scholars in the humanities often create and work with electronic texts, and have come to appreciate many of the possibilities inherent in machine-readable text. But the impact of these possibilities has largely been restricted to what I would call the "procedural" side of the humanist's labor:

- The mutability of electronic texts (e-texts) is utilized only in editing and recycling texts.
- The reproducibility of machine-readable texts is used mainly to transform them into print format.
- The ease of transmission of e-texts across electronic networks is used for only minimal dissemination.
- The wide array of search possibilities has been limited mainly to citation and lookup queries.

This reductionist, procedural, model has retarded the use of machine-readable texts in the "creative" or "intellectual" realm -- the humanist's ultimate domain of textual analysis and interpretation.

Of course, a computer cannot process "information" that is not explicitly considered or excluded -- unless it refers to an electronic thesaurus, it cannot recognize the word *beautiful* as a synonym of the word *lovely*. Secondly and more importantly, computers have little tolerance for ambiguity. The subtleties of a concept, a metaphor, or an image need to be translated into the rigid confinement of a one-dimensional word. In the area of computer-aided textual analysis, this apparent inability to deal with the implicit and the ambiguous has prevented some of the distinct advantages of computing from permeating into the humanist's intellectual realm.

Much text-directed computer research has been marked by these inabilities -- often negatively, resulting in quantitative number-crunching and statistical analyses, but sometimes positively, using the computer's computational strengths to facilitate or underpin

interpretative statements. Michael Riffaterre studied the repetition of verbal mannerisms in Gobineau's *Pléiades*, such as *assurément*, *sans doute*, and *soit* (*Le Style des Pléiades de Gobineau*, New York, 1957). His conclusions, made possible and supported by computerized stylistic analysis, went beyond statistical banalities and recognized Gobineau's tic words as "linguistics tips of a psychological iceberg revelatory of the deep currents of thoughts, of the major ideas, and even of the *idées fixes* of the author, [which] can help us understand his conscious choices." Another worthwhile piece of research is the authorship study of Mosteller and Wallace on the disputed papers in *The Federalist* (*Inference and Disputed Authorship: The Federalist*, Reading, 1964). In this study, the occurrence of specific words in the attributed papers and in the twelve disputed papers helped lead to the conclusion that Madison wrote the disputed papers.

But the overall record, after almost 40 years of text-directed computing, is rather disappointing. Many of the past studies have failed "to produce results of sufficient interest, rigor and appeal to attract a following among scholars who *do not* make extensive use of computers." There are three reasons for this. First of all, the problems of ambiguity and explicitness have not been adequately resolved. Secondly, many humanities-computing specialists have failed to focus on how text-directed computing could help the analytical and interpretive process, or even might change the type of questions a scholar would ask from the text -- in other words, a lot of humanities computing has been done for its own sake. And finally, humanities-computing specialists have not developed their own theoretical framework, nor tried to link the new possibilities of text-directed computing to some of the theoretical concerns of humanities scholars.

Recent discussion and evolutions in the areas of text-directed computing and the encoding of texts might offer us indications in resolving these problems.

Mark Olsen has proposed a way to link the strengths of text-directed computing with theoretical concerns. In the Humanist electronic conference (Footnote 1), Olsen wrote, "The corrective is to engage and exploit the developments in critical theory head on. Indeed, it is my firm belief that the technology allows us to rethink the notion of 'textuality' and the relationship of text to context (discursive, social, and political). And provide solid, verifiable results based on new theoretical models, allowing us to test and (hopefully) improve critical theory. Humanities computing should be in the lead of rethinking textuality precisely because the technology allows us to treat text as a

radically different object of research." One way to derive more convincing results might be to study issues like intertextuality, through the analysis of a broad body of texts, rather than to concentrate on the individual text. Olsen pointed out the failure of computer-aided literature studies "results from past concentration on in-depth studies of individual texts or authors, studies seeking to identify subtle semantic or grammatical structures, precisely the areas in which computer processing is the weakest."

Certain databases like ARTFL (American and French Research on the Treasury of the French Language), which can provide simultaneous access to a large corpus of texts, could be instruments for this type of research. Consider the work of Keith Baker: for his book *Inventing the French Revolution* (Cambridge, 1990), he used the ARTFL database to study the idea of "public opinion." ARTFL, he states, "was enormously useful in identifying occurrences of *opinion publique* in the database for further analysis, in suggesting a tentative chronology for the usage of the term in eighteenth-century France, and in illustrating the traditional associations of *opinion* with uncertainty, instability, and disorder -- associations that were rapidly changed when mere *opinion* was transformed (as it was during the third quarter of the eighteenth century) into the rational authority of *opinion publique*, the new tribunal to which all political actors were compelled to appeal."

To claim a preference for analyzing a corpus of texts while abandoning the individual text might be an interesting tactical retreat, and it could produce worthwhile results and partially fill the present theoretical void. On the other hand, it evades the pervasive problem of ambiguity and explicitness, which becomes most apparent in the computer-aided analysis of individual texts. Attempts to analyze individual texts rigorously through ARTFL will not be very successful, since the database does not accommodate the separation of text from search program, thus restricting the analysis to the limited possibilities of the program and making it quite impossible for scholars to manipulate the text in meaningful ways (for instance, by incorporating data that could increase the searchability of the text and yield more complex output -- additions known as "markup," discussed below). Even a new version of *PhiloLogic* (the search program for ARTFL) will not really improve matters in this area. Nonetheless, to abandon the individual text as a legitimate object for computer-aided analysis would be a costly capitulation -- could cause humanists to retreat entirely from text-directed computing.

Recent developments in the area of markup or encoding could prove to be very fruitful for computer-aided textual analysis of both single texts and bodies of texts. Markup is the addition of extratextual elements to an electronic text. A formalized markup language provides conventions that identify markup, regulate its usage, and allow it to be distinguished from the text itself. In the area of text-directed computing, markup could, to some degree, compensate for the inadequacies of computers. Willard McCarty distinguishes two types of text directed computing: "Blind" computing is an exploration of a text without significant input of the user's knowledge and ideas about it -- in other words, an algorithmic approach such as ARTFL. "Catoptric" computing, in contrast, is the close, recursive examination of a text that is significantly and increasingly enriched by the user's ideas -- in other words, a metatextual approach which requires tagging or markup. Markup would allow the scholar to enrich the text with information that could be used to analyze the text with greater subtlety and ambiguity. Tagging would be a process of the human mind (I tag, therefore I think). This would not, however, completely solve the problem of "disambiguation," since distinctions and choices would still need to be made; but these would be human factors that typify thinking, and not a computer-driven compulsion to disambiguate.

Past encoding schemes and markup languages have often reflected the research interests of their originators, confined to only one subject area and one applications program. This diversity in encoding schemes, of course, prevents the diachronic type of textual research described above, since the encodings of different e-texts would be incompatible. Since 1987, however, the Text Encoding Initiative (Footnote 2) has adopted a common interchange standard for encoding machine-readable texts: the Standard Generalized Markup Language (SGML, set forth as ISO 8879). Within the syntactic framework of SGML, an encoding scheme can be designed that can handle all the intellectual problems a scholar might want to encode. Furthermore, SGML's Document Type Definition (DTD) allows for the internal validation and consistency of the encoding tags; the TEI header that precedes the SGML encoded transcription of the text satisfies the need of the scholar to have exact information about the data, its source, and its markup. Specifics are included on

- The file description: title of the file; the funding sources; names of those who captured, encoded, and validated the e-text.
- The source that was used to create the e-text: bibliographic information such as author, title, editor, and imprint.

- The encoding particulars.
- Any revisions by the original encoder or succeeding encoders of the e-text.

The groundwork that has been laid with the TEI will allow scholars like Mark Olsen to explore issues of intertextuality in a corpus of multiple texts, as it allows Willard McCarty to proceed with his "catoptric" analysis -- which could become a cumulative and collaborative effort. The TEI will prove to be essential for the production of high-quality electronic texts that will withstand scholarly scrutiny. Since many electronic texts are created, directly or indirectly, for the commercial market, it is of vital importance that scholars and librarians demand that commercial considerations not dilute the high standards of textual and critical editing that can be provided through TEI-conformant electronic texts.

Scholars will then be able to enjoy all the advantages of a machine-readable text:

- The mutability of electronic texts will allow scholars to manipulate, revise, encode, and edit texts that will become instruments to advance and underpin their own textual analysis and interpretation; but others can also use these texts to verify research results, or to challenge and change interpretations by providing alternative tagging.
- The reproducibility of machine-readable texts will allow them to be transformed, preserved, and used in future media.
- The ease of transmission of e-texts across electronic networks will allow for alternatives to the current system of publication.
- Ultimately, the wide array of search possibilities will allow new questions to be asked, new ways to envision texts.

In other words, a tool is not just a tool; it has the potential to revolutionize our perspective. Electronic texts will do just that.

To fulfill this goal, we need a larger body of TEI-conformant SGML encoded e-texts, both new texts, edited by scholars and published by individuals, scholarly societies and commercial publishers, and older texts scanned with improved OCR (optical character recognition) equipment and encoded according to the TEI Guidelines. Libraries will have to collect and archive electronic texts and provide local and remote access to them. Technologically and philosophically, research libraries are in a good position to cope with e-texts and support the scholar's future research requirements; but they will need to address

certain budget issues if they are to meet these challenges effectively. The development of software for text analysis has to be encouraged, but it must be both easier to use and independent of specific programming language. Specialists in humanities computing need to emphasize the *intellectual nature* of their work; we can then look forward to more exciting and meaningful research that incorporates text-directed computing as one of its tools.

Footnotes

1. The quotations from Mark Olsen appeared in his papers to the *Humanist* listserv in volume 6, number 0364 (Nov. 20, 1992), and vol 6, number 0385 (Dec. 8, 1992).
2. TEI (Text Encoding Initiative) is a joint project of the Association for Computers and the Humanities (ACH), the Association for Computational Linguistics (ACL) and the Association for Literary and Linguistic Computing (ALLC). The project is funded by the National Endowment for the Humanities, the Commission of the European Communities (Directorate XIII), the Andrew W. Mellon Foundation and the Social Science and Humanities Research Council of Canada. The TEI advisory board is made up of representatives of 15 scholarly organizations. The editors of the TEI Guidelines are Lou Burnard (Oxford U.) and C.M. Sperberg-McQueen (U. of Illinois at Chicago). Informed and technical discussion of the TEI takes place on the electronic discussion list TEI-L. For further discussion of the TEI and SGML see Burnard & Sperberg-McQueen "Living with the Guidelines: An Introduction to TEI tagging," Aug. 16, 1992.

Further Reading

On the use of ARTFL in research and the classroom, see *The ARTFL Project Newsletter*, vol. 8, no. 1, winter 1992-93, which includes Keith Baker, "Public Opinions and Revolutionary Thoughts: Searching for Eighteenth-Century Political Culture," cited in this article.

On electronic critical editions and the TEI, see Charles B. Faulhaber, "Textual Criticism in the 21st Century," in *Romance Philology*, vol. 45, no. 1, Aug. 1991, pp. 123-48.

On electronic publishing and changes in scholarly communication, see Anthony M. Cumming, et al., *University Libraries and Scholarly Communication: A Study prepared for the Andrew W. Mellon*

Foundation. Washington, D.C., The Association of Research Libraries, Nov. 1992, especially pp. 123-39.

On the development of software for textual analysis, see Nancy Ide and Jean Veronis, "What Next, After the Text Encoding Initiative? The Need for Text Software," in *ACH Newsletter*, winter 1993.