SUSAN HOCKEY

Center for Electronic Texts in the Humanities
Rutgers and Princeton Universities
New Brunswick, New Jersey

Electronic Texts in the Humanities: A Coming of Age

ABSTRACT

Electronic texts have been used for research and teaching in the humanities ever since the end of the 1940s. This paper charts the development of various applications in literary computing including concordances, text retrieval, stylistic studies, scholarly editing, and metrical analyses. Many electronic texts now exist as a by-product of these activities. Efforts to use these texts for new applications led to the need for a common encoding scheme, which has now been developed in the form of the Text Encoding Initiative's implementation of the Standard Generalized Markup Language (SGML), and to the need for commonly used procedures for documenting electronic texts, which are just beginning to emerge. The need to separate data from software is now better understood, and the variety of CD-ROM-based text and software packages currently available is posing significant problems of support for libraries as well as delivering only partial solutions to many scholarly requirements. Attention is now turning to research towards more advanced network-based delivery mechanisms.

INTRODUCTION

It is now forty-five years since Father Roberto Busa started work on the first-ever humanities electronic text project to compile a concordance to the works of St. Thomas Aquinas and related authors (Busa 1974-). Since that time, many other electronic text projects have

begun, and a body of knowledge and expertise has gradually evolved. Many lessons have been learned from these activities, and it is now possible to make some realistic projections for the future development of electronic text usage in the humanities. Until recently, almost all work has been done on electronic transcriptions of text rather than on digitized images. The discussion in this paper will concentrate on transcriptions, which are referred to as text, but the implications for images will be noted briefly.

The focus of the paper is on primary source material in the humanities. This can be literary text, which is prose, verse, or drama, or a combination of these. It may also be documentary and take the form of letters, memoranda, charters, transcripts of speeches, papyri, inscriptions, newspapers, and the like. Other texts are studied for linguistic purposes, notably collections of text forming language corpora and early dictionaries. Many humanities texts are complex in nature, and the interpretation of the complex features within them is often the subject of scholarly debate. Some texts contain several natural languages and/or writing systems. Others have variant spellings, critical apparatus with variant readings, marginalia, editorial emendations, and annotations, as well as complex and sometimes parallel canonical referencing schemes. An adequate representation of these features is needed for scholarly analysis.

APPLICATIONS IN LITERARY COMPUTING

The earliest and most obvious application was the production of printed word indexes and concordances, often with associated frequency lists. A word index is a list of words in a text where each word (keyword) is accompanied by a reference indicating the location of the occurrences of that word in the text. In a concordance, each occurrence of each word is also accompanied by some surrounding context, which may be a few words or up to several lines. A word frequency list shows the number of times that each word occurs. Words would normally appear in alphabetical order, but they could also be alphabetized or sorted by their endings, which is useful for the study of morphology or rhyme schemes, or in frequency order where the most common words or the *hapax legomena* (once-occurring words) can easily be seen. Specialized concordances show words listed by their references (for example, by speaker within a play) or sorted according to the words before or after the keyword, or by the number of letters they contain. It can be seen that the production of concordances was typically a mechanical batch process that could generate vast amounts of printout.

Early on, attention was also paid to defining the alphabetical order for sorting words in a variety of languages, for example, transcriptions of Greek and Russian as well as Spanish where *ch, ll,* and *rr* are separate

letters of the alphabet. Ways of dealing with hyphens, apostrophes, accented characters, editorial emendations, and the like were soon devised, and in most cases, the choice was left to the user. A major strength of two of the most widely used concordance and retrieval programs today, Micro-OCP and TACT, is their flexibility in alphabet definitions. More detail on alphabetization and different types of concordances may be found in Howard-Hill (1979), Hockey (1980), and Sinclair (1991).

By the mid-1950s, a number of other concordance-based projects had begun. Brandwood's (1956) work on Plato formed the basis of a stylistic study. In France, plans for the Trésor de la Langue Française, a vast collection of literary works since the time of the revolution, began in 1959 to aid the production of the new French dictionary (Quémada 1959). These texts form the basis of the ARTFL (American Research on the Treasury of the French Language) database at the University of Chicago. Other groups or projects of note in the 1960s include Howard-Hill's (1969) Oxford Shakespeare Concordances, word frequency counts of Swedish (Gothenburg) (Allén 1970), Classical Latin texts at Liège (Delatte and Évrard 1961), Medieval Latin in Louvain-la-Neuve (Tombeur 1973), and work on various Italian texts at Pisa under the direction of Antonio Zampolli (1973). At that time, the only means of input was uppercase-only punched cards or, sometimes, paper tape. Burton (1981a, 1981b, 1981c, 1982) describes these projects and others in her history of concordance making from Father Busa until the 1970s, which makes interesting reading.

The interactive text retrieval programs that we use today are a derivative of concordances, since what they actually search is a precompiled index or concordance of the text. Besides their obvious application as a reference tool, concordance and text retrieval programs can be used for a variety of scholarly applications, one of the earliest of which was the study of style and the investigation of disputed authorship. The mechanical study of style pre-dates computers by a long time. Articles by T. C. Mendenhall at the end of the last century describe his investigations into the style of Shakespeare, Bacon, Marlowe, and many other authors, using what seems to have been the first-ever word-counting machine. Mendenhall (1901, 101-2) notes

> the excellent and entirely satisfactory manner in which the heavy task of counting was performed by the [two] ladies who undertook it. . . . The operation of counting was greatly facilitated by the construction of a simple counting machine by which the registration of a word of any given number of letters was made by touching a button marked with that number.

Mendenhall's findings were not without interest, since he discovered that Shakespeare has more words of four letters than any other length, whereas almost all other authors peak at three. Many other stylistic

studies have based their investigations on the usage of common words, or function words. These are independent of content, and authors often use them unconsciously. Synonyms have also been studied as have collocations or pairs of words occurring close together. The work of Mosteller and Wallace (1964) on the *Federalist Papers* is generally considered to be a classic authorship study, since the twelve disputed papers were known by external evidence to be either by Hamilton or by Madison and there was also a lot of other material of known authorship (Hamilton or Madison) on the same subject matter. A study of common words showed that Hamilton prefers "while," whereas Madison almost always uses "whilst." Other words favored by one or the other of them included "enough" and "upon."

Anthony Kenny's (1978) investigation of the *Aristotelian Ethics* was based on function words, which he divided into categories such as particles and prepositions, that were derived from his reading of printed concordances. He was able to show that the usage of common words in three books that appear in both the *Nicomachean* and the *Eudemian Ethics* is closer to *Eudemian Ethics*. More recently, John Burrows's (1987) examination of Jane Austen's novels has become something of a landmark study in literary computing. By analyzing their usage of common words, he was able to show gender differences in the characters in the novels and to characterize their idiolects. These and similar studies employ some simple statistical methodologies for which Kenny (1982) is a useful introduction. They also show the need to index every word in the text and to distinguish between homographic forms.

Concordances can also be a valuable tool for the historical lexicographer, and several large textbases were originally compiled for this purpose. The Dictionary of Old English (DOE) in Toronto created the complete Corpus of Old English, which totals some three million words. Lexicographers at the DOE have created complete concordances of all this corpus and select citations from the concordances for the dictionary entries (Healey 1989). The most frequent word in Old English occurs about 15,000 times, and it was just possible for a lexicographer to read all the concordance entries for it. This is obviously not feasible for much larger corpora such as the Trésor de la Langue Française. A notable modern example of what has become known as corpus lexicography is Collins's *COBUILD English Dictionary,* which was compiled using a twenty-million-word corpus of English (Sinclair 1987).

Other electronic texts have been created for the analysis of meter and rhyme schemes. In the 1960s, scansion programs existed for Greek and Latin hexameter verse (Ott 1973). Metrical dictionaries were compiled for authors as diverse as Hopkins (Dilligan and Bender 1973) and Euripides (Philippides 1981). Sound patterns have been studied in Homer

(Packard 1974), some German poets (Chisholm 1981), and Dante (Robey 1987).

The traditional scholarly editing process has also led to the creation of some electronic texts. In simple terms, this process has consisted of collating the manuscripts, establishing the textual tradition, compiling an authoritative text, compiling the critical apparatus, and then printing text. In the 1960s, computer programs to collate manuscripts began to appear, and it was soon realized that collation could not be treated as a completely automatic process and that, because of the lineation, verse was easier to deal with than prose. Robinson's (forthcoming) COLLATE program was developed after a study of earlier systems. It has a graphical user interface and is by far the most flexible collation program.

Many early humanities projects were hampered by design forced upon them by the limitations of hardware and software. Until disk storage became more widely available in the 1970s, texts and associated material were stored on magnetic tape, which could only be accessed sequentially. Disk storage allowed random access, but data were still constrained within the structures of database programs, particularly relational databases where the information is stored as a set of rectangular tables and is viewed as such by the user. Very little humanities-oriented information fits this format without some restructuring, which, more often than not, results in some loss of information.

Hypertext has provided a solution to data modeling for the humanities. It offers flexible data structures and provides a web of interrelated information, which can be annotated by the user if desired. An obvious application in the humanities is the presentation of primary and secondary material together. Images, sound, and video can be incorporated to aid the interpretation of the text. The traditional scholarly edition can be represented very effectively as a hypertext, but hypertext is a more obvious medium for presenting multiple versions of a text without privileging any particular one of them (Bornstein 1993). Other experiments have used hypertext to model the narrative structure of literature with a view to helping students understand it better (Sutherland forthcoming).

ELECTRONIC TEXTS TODAY

Many of the electronic texts that are in existence today were created as a by-product of research projects such as those described above. Large collections of text have been assembled by a few research institutes, mostly in Europe where public money has been provided for the study of language and its relation to the cultural heritage. Most other texts

have been compiled by individuals for their own projects. These texts reflect the interests of those research groups or individuals, and it is perhaps questionable as to how many of them can be used for other scholarly purposes. These texts are ASCII files, not files that have been indexed for use by specific programs. Initial estimates show that 90 to 95 percent of texts fall into this category. For a variety of reasons, few of them have been made available for other scholars to use, and these scholars may find that they are not well suited to their purposes.

However, it was soon realized that considerable time and effort is required to create a good electronic text. Many existing texts have been keyboarded, and this is still the normal means of input. Optical character recognition (OCR) of some material became feasible in the early 1980s, but in general, it is only suitable for modern printed material. OCR systems tend to have difficulty with material printed before the end of the last century, newspapers, or anything else where the paper causes the ink to bleed, as well as material containing footnotes and marginalia, nonstandard characters and words in italic, or small capitals. Those systems that are trainable can be more suitable for humanities material, but these require some skill on the part of the operator. Hockey (1986) and the collection of papers assembled by the Netherlands Historical Data Archive (1993) give further information. More importantly, OCR also generates only a typographic representation or markup of the text, whereas experience with using texts has shown that this is inadequate for most kinds of processing and analysis. Most large data entry projects are choosing to have their data keyed, which allows some markup to be inserted at that time.

Recognizing the need to preserve electronic texts, the Oxford Text Archive (OTA) was established in 1976 to "offer scholars long term storage and maintenance of their electronic archives free of charge." It has amassed a large collection of electronic texts in many different formats and is committed to maintaining them on behalf of their depositors. Depending on the conditions determined by their depositors, OTA texts are made available to other individuals for research and teaching purposes at little cost. However, there is no guarantee of accuracy, and users of OTA texts are encouraged to send any updated versions that they may have created back to Oxford. Proud (1989) reports on the findings of a British Library sponsored project to review the Oxford Text Archive.

There have been a few systematic attempts to create or collect and archive texts for general-purpose scholarly use. The most notable one for a specific language is the Thesaurus Linguae Graecae (TLG), which began at Irvine, California, in 1972. It is now nearing completion of a databank of almost seventy million words of Classical Greek (Brunner 1991). The texts are distributed on a CD-ROM that contains plain ASCII

files. They are not indexed in any way. In the late 1980s, the Packard Humanities Institute (PHI) compiled a complementary CD-ROM of all Classical Latin, which is about eight million words. The Women Writers' Project at Brown University is building a textbase of women's writing in English from 1330 to 1830 and contains many texts that are not readily accessible elsewhere. Begun in the 1980s, the Dartmouth Dante Project (DDP) is aiming to make available the text of the *Divine Comedy* and all major commentaries. The texts are stored and indexed using BRS-Search and can be accessed via Telnet to lib.dartmouth.edu then, at the prompt, type "connect dante."

A few other collections of text should be noted here. The Istituto di Linguistica Computazionale in Pisa has a large collection of literary and nonliterary works in Italian. Institutes funded by the German government at Bonn and Mannheim have been building text collections for many years. Bar-Ilan University in Israel is the home of the Responsa Project, and the Hebrew Academy in Jerusalem also has a substantial collection. Material in Welsh and other Celtic languages has been built up at Aberystwyth and elsewhere. The International Computer Archive of Modern English at Oslo concentrates on English-language corpora, and groups in various English-speaking countries are compiling corpora of their own usage of English. The British National Corpus is nearing completion of a hundred-million-word corpus of written and spoken English. Many other similar activities exist. The Georgetown University Center for Text and Technology maintains a catalog of projects and institutes that hold electronic texts but not the texts themselves. This catalog can be accessed most easily by Gopher to guvax.georgetown.edu. Lancashire (1991) is the most comprehensive source of information in print about humanities computing projects in general.

The Rutgers Inventory of Machine-Readable Texts in the Humanities is the only attempt to catalog existing electronic texts using standard bibliographic procedures (Hoogcarspel 1994). The Inventory is held on the Research Libraries Information Network (RLIN) and is maintained by the Center for Electronic Texts in the Humanities (CETH). It contains entries for many of the texts in the Oxford Text Archive, plus material from a number of other sources. The Inventory is now being developed by CETH staff who have prepared extensive guidelines for cataloging monographic electronic text files using *Anglo-American Cataloguing Rules,* 2d ed., (*AACR2*) and RLIN.

In the last few years, more electronic texts have begun to be made available by publishers or software vendors. These are the texts that are more likely now to be found in libraries. They are mostly CD-ROMs and are usually packaged with specific retrieval software. Examples include the *Global Jewish Database* on CD-ROM, the *New Oxford*

English Dictionary on CD-ROM, the CETEDOC CD-ROM of the Early Christian Fathers, and the WordCruncher disk of American literature. The CD-ROM versions of the *English Poetry Full-Text Database* and *Patrologia Latina* published by Chadwyck-Healey also fall into this category, although these texts are also available on magnetic tape for use with other software. Oxford University Press also publishes electronic texts, which are ASCII files. Their texts are particularly well documented, and most can be used with the Micro-OCP concordance program, which they also publish.

Some of these packaged products are relatively easy to use, but prospective purchasers might want to be aware of a number of issues before they launch into acquiring many of them. Almost every one of these products has its own user interface and query language. They are mostly designed for scholarly applications on what are complex texts. Therefore, it can take some time to understand their capabilities and to learn how to use them. If this proliferation of products continues, the cost of supporting them will not be insignificant. Librarians are not normally expected to show patrons how to read books, but they can expect to spend some considerable time in learning how to use these resources and showing them to users. Those that are easy to use may not satisfy many scholarly requirements. For example, on the WordCruncher CD-ROM, which is one of the easiest to use, the texts have been indexed in such a simple way that there is no way to distinguish between I in act and scene numbers (e.g., Act I) and the pronoun I. Several of these products are designed for the individual scholar to use on his or her own machine rather than for access by many people. They provide good facilities for storing search requests for future use, but this is not much help if twenty other people have stored new requests or modified existing ones in between. Another issue is just what words have been indexed and how. A response to any search request is only as good as the words that have been indexed. In some cases, this seems to have been determined by software developers who have little understanding of the nature of the material and the purposes for which it might be used. Other institutions have chosen to acquire texts in ASCII format and provide network access to them, usually with Open Text's PAT system. In this case, the burden of deciding what to index falls on the librarian, who is thus assuming some responsibility for the intellectual content of the material.

CREATING ELECTRONIC TEXTS FOR THE FUTURE

Creating an electronic text is a time-consuming and expensive process, and it therefore makes sense to invest for the future when doing

it. Texts that are created specifically for one software program often cannot easily be used with others. The need to separate data from software is now well recognized. Data that are kept in an archival form independent of any hardware and software stand a much better chance of lasting for a long time because they can be moved from one system to another and because they can be used for different purposes and applications.

Experience has shown that an archival text needs markup and documentation for it to be of any use in the future. Markup makes explicit for computer processing things that are implicit to the human reader of a text. Markup is needed to identify the structural components of a text (chapter, stanza, act, scene, title) and enables specific areas or subsets of text to be searched and text that has been retrieved to be identified by references or other locators. It may also be used to encode analytic and interpretive features. Many humanities texts are complex in nature, and many different markup schemes have been created to encode their properties. Ones that have been in common use are COCOA, which is used by Micro-OCP and TACT, the beta code used by the Thesaurus Linguae Graecae, and the three-level referencing system used by WordCruncher. These markup schemes concentrate on the structure of a text, as opposed to schemes such as TeX and troff, which contain formatting instructions.

Following a planning meeting in 1987, a major international effort to create guidelines for encoding electronic texts was launched by the Association for Computers and the Humanities, the Association for Computational Linguistics, and the Association for Literary and Linguistic Computing. This project, known as the Text Encoding Initiative (TEI), brought together groups of scholars, librarians, and computer professionals to examine many different types of texts and to compile a comprehensive list of the features within those texts.

The TEI soon determined that the Standard Generalized Markup Language (SGML) was a sound basis for the development of the new encoding scheme. SGML became an international standard in 1986. It is a metalanguage within which encoding schemes can be defined. It is descriptive rather than prescriptive and thus can form the basis of the reusable text. It permits multiple and possibly conflicting views to be encoded within the same text. It is incremental so that new encodings can be added to a text without detriment to what is already there. SGML-encoded texts are also ASCII files, and so their longevity can be assured. The TEI's application of SGML is very wide ranging. It provides base tag sets for prose, verse, drama, dictionaries, transcripts of speech, and terminological data. To these can be added tag sets for textual criticism, transcription of primary sources, language corpora, formulae and tables, graphics, hypermedia, analytical tools, and names

and dates. The application has been designed so that other tag sets can be added later. The first definitive version of the TEI guidelines has very recently been published (Sperberg-McQueen and Burnard 1994).

Many existing electronic texts have little or no documentation associated with them. Often, it is difficult to establish what the text is, where it came from, and, in a few cases, even what language it is in. There seem to be two main reasons for this. In some cases, the text was created by an individual who was so familiar with that text that he or she did not find it necessary to record any documentation about it. In other cases, the person who created the text did not have any model to follow for documenting the text and thus recorded only minimal information about it. Where documentation does exist, it is in many different formats, making the task of compiling information about electronic texts extremely difficult.

As part of its recommendations, the TEI has proposed an electronic text file header to meet the needs of librarians who will manage the texts, scholars who will use them, and computer software developers who will write programs to operate on them. The TEI header consists of a set of SGML elements that include bibliographic details of the electronic text and the source from which it was taken, information about the principles that governed the encoding of the text, any classificatory material, and a revision history that records the changes made to the text.

DIGITAL IMAGING

Many of the lessons learned from the creation and use of electronic texts can also be applied to digital imaging of manuscripts and textual material. The potential of digital imaging for preservation and access is now being exploited in numerous projects. From this point of view, the archival role is obviously very important. Most of the cost in digital imaging is in taking the object to and from the camera, and so it makes sense to digitize at the highest resolution possible. Storing the image in a proprietary format linked to some specific software will lead to all the same problems that have been experienced with text stored in a proprietary indexing program. It will not be possible to guarantee that the image will be accessible in the future or that it can be used for other purposes. Documentation and provenance information are just as important for images. SGML can be used to describe material that is not itself textual. The TEI header would need only a slight modification to be used for images and offers a route to using both text and image together. The TEI's hypertext mechanisms allow pointers from the text to the image and can form the basis of a system that operates on the transcription of the text but displays the image to the user.

ANALYSIS TOOLS

Experience of working with electronic literary texts has highlighted a number of analysis tools and features that have been found to be useful. The most obvious is the need to index every word and not to have a stop list. This is important for many stylistic and linguistic studies that have concentrated on the usage of common words. It also avoids the omission of some homographic forms; for example, the English auxiliary verbs "will" and "might" are also nouns. The punctuation is often important in early printed texts, and some scholars may want to search on that. In other languages, it provides a simple key to the examination of the ends of sentences, for example, clausulae in Classical Latin. Words that are not in the main language of the texts need to be indexed separately to avoid homographs such as "font" in English and French, or "canes" in English and Latin. The ability to search on the ends of words is also useful, particularly for verse and in languages that inflect heavily. A very small number of resources provide an index by endings. For others, this kind of search can take some time as it can only be handled by a sequential search on the word index. A good text will also have structural encoding, and the user may want to have the option of restricting proximity searches to within certain structural boundaries or allowing them to extend beyond a boundary. For example, finding "tree" within ten words of "flower" may not be useful if "tree" is the last word of a chapter and "flower" occurs at the beginning of the next chapter.

There has not been as much progress in the development of tools to analyze text. Essentially, we are still able to search text only by specifying strings of characters, possibly linked by Boolean operators, whereas most users are interested in concepts, themes, and the like. String searches cannot effectively disambiguate homographic forms, for example, "bank" as in money bank as opposed to "bank" of the river or the verb "bank" (used of an airplane), or Latin "canes" as "dogs" or "you will sing."

Computer programs to perform morphological analysis, lemmatization, syntactic analysis, and parsing have been used experimentally for some time, but our understanding of these is still only partial. The most successful parsing programs claim accuracy of about 95 percent. Morphological analysis has been done reasonably well for some languages, for example, Ancient Greek, but there are no widely available general-purpose programs that are suitable for literature. Father Busa recognized the need to lemmatize his concordance to St. Thomas Aquinas in order to make it more useful to scholars, but this was done manually, which is still the only way to ensure accurate data.

In Busa 1992, he reflects on the lack of intellectual progress and on how little the computer can still do.

Because of its nature, literature is harder to deal with than many other types of text, and there have been relatively few attempts to apply more sophisticated language analysis algorithms to it. After years of working with rule-based systems, researchers in computational linguistics are turning to the compilation of large-scale lexical resources and knowledge bases for use by natural language understanding systems. The usual method has been to create an electronic version of a printed dictionary and restructure that within the computer as a lexical database that contains morphological analyses, lemmas, frequent collocations, and other information that would help to disambiguate homographic words. However, printed dictionaries are designed for humans not for computers to use. They exist to document the language and thus contain many citations for uncommon usages of words but very few (in proportion to their occurrences) of usual usages. A computer program must look every word up in the dictionary and thus needs more information about common words. This has led to the current interest in language corpora, which are large bodies of text from which information can be derived to augment electronic dictionaries. In many ways, this development represents another coming of age, since the initial methodologies used by computational linguists to analyze large corpora are concordance based and are very similar to those that have been used in literary computing for many years. Once information about word usage has been derived, it can be encoded within the text (using SGML markup) and used to train and refine future programs, which will eventually perform more accurate analyses. We can only hope that this coming of age will lead to better access technologies and to the computer doing more for us.

REFERENCES

Allén, Sture. 1970. Vocabulary Data Processing. In *Proceedings of the International Conference on Nordic* and *General Linguistics,* ed. Hreinn Benediktsson, 235-61. Reykjavik: Visindafelag Islendiga.

Bornstein, George. 1993. What is the Text of a Poem by Yeats? In *Palimpsest: Editorial Theory in the Humanities,* ed. George Bornstein and Ralph G. Williams, 167-93. Ann Arbor: University of Michigan Press.

Brandwood, Leonard. 1956. Analysing Plato's Style with an Electronic Computer. *Bulletin of the Institute of Classical Studies 3*: 45-54.

Brunner, Theodore F. 1991. The Thesaurus Linguae Graecae: Classics and the Computer. *Library Hi Tech 9*(1): 61-67.

Burrows, J. F. 1987. *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method.* Oxford: Clarendon Press.

Burton, Dolores M. 1981a. Automated Concordances and Word Indexes: The Fifties. *Computers and the Humanities 15*(1): 1-14.

Burton, Dolores M. 1981b. Automated Concordances and Word Indexes: The Early Sixties and the Early Centers. *Computers and the Humanities 15*(2): 83-100.

Burton, Dolores M. 1981c. Automated Concordances and Word Indexes: The Process, the Programs, and the Products. *Computers and the Humanities 15*(3): 139-54.

Burton, Dolores M. 1982. Automated Concordances and Word-Indexes: Machine Decisions and Editorial Revisions. *Computers and the Humanities 16*(4): 195-218.

Busa, Roberto, S. J. 1974-. *Index Thomisticus.* Stuttgart-Bad, Connstatt: Fromann-Holzboog.

Busa, Roberto, S. J. 1992. Half a Century of Literary Computing: Towards a 'New' Philology. (Reports from Colloquia at Tübingen). *Literary & Linguistic Computing 7*(1): 69-73.

Chisholm, David. 1981. Prosodic Approaches to Twentieth-Century Verse. *ALLC Journal 2*(1): 34-40.

Delatte, L., and É. Évrard. 1961. Un Laboratoire d'Analyse Statistique des Langues Anciennes à l'Université de Liège. *L'Antiquité Classique 30:* 429-44.

Dilligan, R. J., and T. K. Bender. 1973. The Lapses of Time: A Computer-Assisted Investigation of English Prosody. In *The Computer and Literary Studies,* ed. A. J. Aitken, Richard W. Bailey, and N. Hamilton-Smith, 239-52. Edinburgh: Edinburgh University Press.

Healey, Antoinette diPaolo. 1989. The Corpus of the Dictionary of Old English: Its Delimination, Compilation and Application. Paper presented at the Fifth Annual Conference of the UW Centre for the New Oxford English Dictionary, St. Catherine's College, Oxford, England.

Hockey, Susan. 1980. *A Guide to Computer Applications in the Humanities.* London: Duckworth; Baltimore: Johns Hopkins University Press.

Hockey, Susan. 1986. OCR: The Kurzweil Data Entry Machine. *Literary & Linguistic Computing 1*(2): 63-67.

Hoogcarspel, Annelies. 1994. The Rutgers Inventory of Machine-Readable Texts in the Humanities. *Information Technology and Libraries 13*(1): 27-34.

Howard-Hill, T. H. 1969. The Oxford Old-Spelling Shakespeare Concordances. *Studies in Bibliography* 22: 143-64.

Howard-Hill, T. H. 1979. *Literary Concordances: A Guide to the Preparation of Manual and Computer Concordances.* Oxford: Pergamon Press.

Kenny, Anthony. 1978. *The Aristotelian Ethics: A Study of the Relationship between the Eudemian and Nicomachean Ethics of Aristotle.* Oxford: Oxford University Press.

Kenny, Anthony. 1982. *The Computation of Style.* Oxford: Pergamon Press.

Lancashire, Ian., ed. 1991. *The Humanities Computing Yearbook 1989-90: A Comprehensive Guide to Software and Other Resources.* Oxford: Clarendon Press.

Mendenhall, T. C. 1901. A Mechanical Solution of a Literary Problem. *Popular Science Monthly 60* (December): 97-105.

Mosteller, Frederick, and David L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist.* Reading, Mass.: Addison-Wesley.

Netherlands Historical Data Archive. 1993. *Optical Character Recognition in the Historical Discipline: Proceedings of an International Workgroup.* Netherlands Historical Data Archive, Nijmegen Institute for Cognition and Information.

Ott, Wilhelm. 1973. Metrical Analysis of Latin Hexameter: The Automation of a Philological Research Project. In *Linguistica Matematica e Calcolatori (Atti del Convegno e della Prima Scuola Internazionale, Pisa 1970),* ed. Antonio Zampolli, 379-90. Florence: Leo S. Olschki.

Packard, David W. 1974. Sound-Patterns in Homer. *Transactions of the American Philological Association 104:* 239-60.

Philippides, Dia Mary L. 1981. *The Iambic Trimeter of Euripides.* New York: Arno Press.

Proud, Judith K. 1989. The Oxford Text Archive. British Library Research and Development Report. London: British Library.

Quémada, Bernard. 1959. La Mécanisation dans les Recherches Lexicologiques. *Cahiers de Lexicologie 1:* 7-46.

Robey, David. 1987. Sound and Sense in the Divine Comedy. *Literary & Linguistic Computing 2*(2): 108-15.

Robinson, Peter M. W. Forthcoming. COLLATE: A Program for Interactive Collation of Large Manuscript Traditions. In *Research in Humanities Computing*, vol. 3, ed. Susan Hockey and Nancy Ide, 32-45. Oxford: Oxford University Press.

Sinclair, John M., ed. 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins.

Sinclair, John M. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sperberg-McQueen, C. M., and Lou Burnard, eds. 1994. *Guidelines for the Encoding and Interchange of Electronic Texts*. Chicago and Oxford: Association for Computers and the Humanities, Association for Computational Linguistics, Association for Literary and Linguistic Computing.

Sutherland, Kathryn. Forthcoming. Waiting for Connections: Hypertexts, Multiplots, and the Engaged Reader. In *Research in Humanities Computing*, vol. 3, ed. Susan Hockey and Nancy Ide, 46-58. Oxford: Oxford University Press.

Tombeur, P. 1973. Research Carried Out at the Centre de Traitement Electronique des Documents of the Catholic University of Louvain. In *The Computer and Literary Studies*, ed. A. J. Aitken, Richard W. Bailey, and N. Hamilton-Smith, 335-40. Edinburgh: Edinburgh University Press.

Zampolli, Antonio. 1973. La Section Linguistique du CNUCE. In *Linguistica Matematica e Calcolatori (Atti del Convegno e della Prima Scuola Internazionale, Pisa 1970)*, ed. Antonio Zampolli, 133-99. Florence: Leo S. Olschki.