A Proposal for
Establishing a Text-processing Facility
at Lawrence University

Submitted to

The Office of Computing Activities

of the National Science Foundation

From

Lawrence University

Appleton, Wisconsin 54911

$41,014

M. O. Wrolstad

Marwin O. Wrolstad
Treasurer and Vice President
  for Business Affairs
Lawrence University
414-739-3681 Ext. 223

Date July 19, 1972

Thomas S. Smith

Thomas S. Smith
President
Lawrence University
414-739-3681 Ext. 220

Date July 19, 1972

Ben R. Schneider, Jr.

Ben R. Schneider, Jr.
Principal Investigator
Professor of English
Lawrence University
  On leave at
2 Gloucester Crescent
London NW1 7DS England
01 485 4514

Date July 19, 1972

ABSTRACT

The interactive Cathode Ray Tube Terminal (CRT), being the best text-processing device ever devised, as well as a form of computer input, greatly simplifies the capture, editing, and analysis of written text. Since the greatest block to computer study of large amounts of text is the high cost of producing machine-readable texts, the installation of interactive CRT text-processing terminals at academic institutions will greatly stimulate computer research in the humanities, facilitate the work of other staff and faculty who have text-processing problems, and generate new forms of research. Lawrence University proposes to demonstrate the truth of this proposition by installing a CRT text-processing terminal in its PDP-11 system and developing text-handling programs for it. The London Stage Information Bank project (LSIB), now being completed at Lawrence, will use this facility for editing its data base and adding new sources of theatrical information. It will also be used for collecting a library of machine-readable literary texts and in faculty research and teaching based on these texts. Since LSIB has taken extensive measures to invite the involvement of computing and literary people in its enterprise, and it has already aroused a great deal of interest among them, there is good reason to believe that LSIB's computing experiences will be shared by a broad spectrum of people having similar problems.

OTHER SOURCES OF FUNDS

This specific proposal is not being submitted to any other agency, but an application for completion of *The London Stage* Information Bank (LSIB); (herewith submitted), in which a sum is budgeted (p. 11) for a Cathode Ray Tube terminal (CRT), has been accepted by the National Endowment for the Humanities on the condition that for every dollar we raise they will give us a dollar. The two proposals differ in that The National Science Foundation is asked to support the *purchase* of a permanent facility, whereas NEH is supporting temporary rental, purchase not being allowed under the terms of the grant. If we can purchase a terminal with NSF help, the NEH budget budget item for a CRT may be cancelled, but if matching funds for the NEH item are collected, we would still rather have a permanent facility in the terms laid down in this proposal and give up the NEH rental funds. Further, if only *some* matching funds are forthcoming, we would rather reserve them for other parts of the budget that NSF could not justify supporting. Various alternatives and possibilities are discussed under *Budget*, pp. 26-27. Both budgets include sums for the development of the facility. If NSF accepts this proposal, we shall subtract the sum requested here from the NEH budget.

Since the project for completion of LSIB, of which I am director, does not officially start until 1 August, I am not at present receiving any concurrent support for these CRT plans, or in fact, for any project. $2500 has been collected to match the NEH grant, and Dean George Winchester Stone, Jr., Chairman of LSIB's advisory board and principal fund-raiser for it, has applied to the Mellon, Billy Rose, Janesville, Bush, Sloan, Harnishfeger, and Ashley Foundations, and the Lilly Endowment. Lawrence also seeks funds.

A PROPOSAL FOR
A TEXT-PROCESSING CATHODE RAY TUBE TERMINAL
AT LAWRENCE UNIVERSITY

It seems evident that the interactive text-editing CRT is a real breakthrough in the preparation of text, possibly the greatest breakthrough since the typewriter. My article on "The Production of Machine-Readable Text" (herewith submitted) gives a full account of what I believe to be its advantages (pp. 39-40; 45-47). In short, not until now has there been a "page" on which appreciable sections can be removed, added, expanded or contracted without necessitating a rewriting, retyping, or resetting of the paragraph, page, or section in order to make the extra space or fill in the space taken out. No writing medium has hitherto provided "elastic space" nor has it been possible to rekey only that part of the text which is in error, without disturbing that part which is already correct. Even a poor typist can produce perfect text simply by correcting the CRT "page" during proofreading.

The CRT is also bound to re-establish the study of written text, especially literary text,on a whole new set of assumptions. The power of mechanical retrieval of specified verbal units, of rapid statistical analysis of repeated verbal phenomena, and the ease of discovering correlations and patterns in verbal material will make possible a solid factual basis for kinds of generalizations about literature that have hitherto been founded more on impressions than evidence. Even when statistical evidence is not presented, literary scholars proceeding along conventional lines but having machine access to the text in question will find evidence to support their statements much faster than is now possible, making possible more work encompassing larger aggregates of text: whole schools of writing, whole periods, complete works, complete literary forms. If, too, most frequently used reference works, catalogues, indexes, bibliographies, encyclopedias, handbooks, companions, etc., are accessible from machines, the scholar's capacity to know will increase by a large factor. Machine-accessible text, in effect, is an extension of the scholar's memory, as the telephone is an extension of his speech and hearing. While the computer increases the scholar's reach and saves his time, it cannot replace him as the imaginative source and rational judge of statements about literary text. The flexibility of approach that is possible with an interactive CRT not only keeps him in command of his medium but also gives scope to his inventiveness.

RESEARCH PURPOSES

More specifically, a CRT text-entering and editing facility at Lawrence University will be used for the following research purposes:

1. Correcting, revising, and indexing the machine-readable text of *The London Stage* as planned for phase two of *The London Stage* Information Bank (*LS*IB) (see Prospectus for Completion, pp. 8-9; "Production of Machine-Readable Text," pp. 3-4, 12-15).
2. Processing new texts to be added to *LS*IB: Bibliographies, Biographical Dictionary of Actors, Texts of plays (Prospectus, pp. 6-7, 10). Since printers' paper tape or magnetic tape produced in the process of printing some of these works may be used for data entry ("Production of Machine-Readable Text,"

p. 45), the terminal will be used to edit this text for computer processing, to enter delimiters, make signposts that trigger program routines fitted to the data structure on hand, and add classificatory tags and labels.

It is expected that the existence of LSIB will encourage altogether new kinds of research in theatre and drama. My own first acquaintance with *The London Stage* ("The Coquette-Prude as an Actress's Line," herewith submitted, see esp. pp. 144-145) stemmed from an attempt to throw light on the meaning of plays by studying casting practices of contemporary London acting companies. I certainly did not begin to exhaust the possibilities of this line of research. A similar quest has led Leonard Leff, one of the scholars who helped LSIB by editing a volume of *The London Stage* for input, in a similar direction. His research plan from his recent paper at the Modern Language Association convention in December 1971, is an example of theatre research already evolving from LSIB. (Report pp. 3, 5).

> Most of us know that The Rivals failed at its first performance. Two cantankerous factions were in attendance, both of them contributing to the chaos. In addition, Shuter was imperfect as Sir Anthony Absolute, Miss Barsanti lisped through Lydia Languish, and Lee hadn't the brogue to play the Irish Sir Lucius O'Trigger. Still, there must have been reasons for casting these and other actors - were they, more than some of their cohorts perhaps, "available"? had they played similar roles before? were they simply late eighteenth century favorites? A computer sort of the principal actors in the premiere will provide answers and should also yield the raw material with which to begin a study of more critical questions. Two examples, first, Laurence Clinch replaced Lee as Sir Lucius in the third performance of The Rivals. A superficial glance at Clinch's career, however, indicates that he played few comic roles. Why was he chosen to replace Lee? Was he known as an especially "quick study", stepping into other premieres or other parts on short notice? Did he specialize in stage Irishmen? And, perhaps incidentally, did his career benefit from "saving" Sheridan's play? A review of Clinch's London career should provide some answers. A second example: the question of Sheridan's sentimentalism promises to endure, yet there may be a resolution to it, at least regarding The Rivals. At its premiere, William T. Lewis performed the critical role of Faulkland, the lovesick youth who serves as Jack Absolute's foil. For decades critics have debated whether Faulkland is the model or the mocker of the sentimental hero. A printout of Lewis' early career will list the roles he was as-

sociated with and the frequency of his appearances in them before coming to Faulkland. Having some indication of the type of character audiences expected from him should be helpful in evaluating what type of character Sheridan originally intended him to play.

The power we have to make very easily a very great number of arrangements and correlations of data in The London Stage will invite a new kind of historical research, possibly providing us with significant views of the forest which the great number of trees has until now kept us from seeing. I imagine that with the help of LSIB we will find ourselves exploring some of the following territory for the first time:

We will be searching for patterns in the data: In what ways is one season like another? In what ways different? What is a typical stage career like? To what extent do actors specialize? What is the effect of the repertory system on actors' careers, casting, play selection, and so forth? Does casting indicate a class structure in stage companies?

We will be looking for trends: the rise of pantomime; the interest in Shakespeare; the proportion of tragedies to comedies; the waning of Restoration comedy; the rise of sentimentalism; the decline of the drama.

Coupling casting information with texts of plays should open more fruitful fields of research: the rise and fall of certain themes; the identification of actors' lines; writing for repertory; an actor's public image, based on the content of characters he plays on the stage.

3. Building a text-library. I plan to teach a course, beginning next year, on the application of computers to humanistic problems, partly theoretical and partly practical. For this course I will want to have a library of machine-readable texts for students to experiment with. One would hope eventually to expand his library to the point where it would be useful for research by students and faculty of the English and foreign

language departments. Student and faculty projects would also contribute to the library's growth.

It has seemed to me that the current practice of publishing by the dozens computer-made concordances of literary works is not the best answer to the scholar's need to retrieve passages of text associated with themes, images, and stylistic traits of a work. Effective research of this sort requires a higher degree of interaction with the text than a concordance allows. The best technique involves planning each step on the results of the previous step. The scholar's approach varies according to the hypothesis to be tested and the nature of the individual literary work in question. The scholar's special knowledge should play a large part in the strategy. Therefore, instead of a concordance, which uses a single very rigid approach to textual retrieval, a scholar should use a machine-readable magnetic tape of the work he wishes to study. Then, using his own institution's computer in consultation with a programmer or a versatile text-handling CRT, or both, he can devise ad hoc routines for the problem in hand. When one considers how simple it can be to call onto a CRT screen all lines or passages in a text that contain a given word, it immediately appears how inefficient a device a concordance is that indexes every word in the text so that you will be sure to find the one word you want (if it wasn't on the "stop list").

4. In addition to these plans involving the London Stage and a text-library, colleagues at Lawrence have plans of their own for computer-aided literary research and teaching that would be difficult to accomplish without a CRT text-processing facility. Herbert Tjossem (English) plans computer study of a large 19th century collection of English dia-

lect words and of variations in vocabulary of Middle English verse from different regions. John and Graciela Alfieri (Spanish) will do a computer study of colloquial language in the novels of Galdós. Peter Fritzell (English) wants to develop a system by which a student can teach himself essential facts about historical backgrounds of American literature by querying a data base. He is interested in the development of a similar system for self-teaching of essentials of poetic analysis, using a number of poetic texts and questions about them. Mojmir Povolny (Government) is doing content analysis of diplomatic exchanges between iron curtain countries and the West. George Smalley (Slavic) has an exhaustive collection of Russian word roots, already keypunched, which offers interesting possibilities for linguistic research with an interactive CRT. Maurice Cunningham (Latin) is engaged in computer study of Latin syntax.

NATIONAL IMPACT AND PUBLICITY

Because LSIB, by virtue of its widely-used subject matter and its experimentation with new methods of creating a data base, reaches out to a broad spectrum of English-speaking scholars and computer people, whatever benefits to the research community accrue from this text-processing venture will probably not be confined to those using the system at Lawrence. LSIB's function as a service to the academic community implies a close relationship with that community, and I have consistently striven to spread the news about its progress for the sake of feedback that might help in its design. During the course of the years 1970 and 1971 announcements of the project appeared in the following journals:

Restoration and 18th Century Theatre Research
Computers and the Humanities
Newsletter of the Special Interest Group in Languages, Social Science and Humanities of the Association for Computing Machinery
Newsletter of the Special Interest Group in Information Retrieval of the Association for Computing Machinery
Publications of the Modern Language Association
Historical Methods Newsletter
Theatre Notebook
Newsletter of the American Educational Theatre Association
The British Studies Monitor
The Johnsonian Newsletter
The Scriblerian
The Appleton, Wisconsin, Post Crescent
The Lawrence University Alumni News

On November 9th 1971 The London Times devoted about six inches to the project as the first item in its "Computer News" department. Shortly thereafter, notices of our project appeared in New Scientist, Computer World, and Computer Weekly. Since LSIB's inception, two news-letters (herewith submitted) reporting our progress and plans have been sent to people who have either participated in the project or who are known to be interested in it. The mailing list has passed 200 and it grows faster all the time.

Notices of the project have elicited 43 enquiries from academic and computing circles and five offers of assistance from commercial information retrieval services.

As part of the endeavour to spread the news, I contributed a paper on text-processing at The Dartmouth Conference on Computers in the Undergraduate Curricula at Hanover in June of 1971. It appeared in the proceedings of that conference and a later version of it was published in Computers and the Humanities, September 1971. Since arriving in England I have participated in a symposium on the subject of a British Theatre Institute and Archive organized by Theatre Quarterly and I have presented to the Society for Theatre Research the case for including in our information bank their bibliography of works devoted to British Theatre from the beginning to 1900. There was a seminar, chaired by Dean George Winchester Stone, Jr. of New York University and attended by 24 persons, at the December meeting of the Modern Language Association, on "The Future and Expansion of *The London Stage* Information Bank".

This spring I gave talks on *LSIB* to the Society for Research in English of London University and to computing and literary people at Glasgow University, and visited literary and computer people at University College, Swansea who are interested in the project. This month I am to speak at Westfield College, University of London, to a Fortran course for people in Arts departments in all colleges of London University and some of the computing staff. At a symposium last March on Computers in Literary Research, sponsored by the Institute for Advanced Studies at Edinburgh University, I gave two papers: one called "Analysis of a Data Base: *The London Stage*," and the other called "Optical Scanning as a Method of Input: The Experience of *The London Stage* Project," both to be published in the proceedings of the conference. Mention of these papers occured in a full-page article on the conference appearing in *The London Times Literary Supplement*. As a result of these papers I was asked to be a consultant for a group of English, Belgian, Swedish and German scholars planning a Dictionary of Early Modern English Pronunciation. I have been asked to give papers next fall at the Midwest Modern Language Association convention and at the annual meeting of the American Society for Information Science.

These public activities are treated at such length to establish that measures have been taken to invite the involvement of computing and literary people and that the project has aroused considerable interest among them. They are listed in response to NSF's stated intention to encourage projects "having a strong national impact" or making "a contribution ... to the specific discipline and to the existing body of computer-based techniques and systems." (Grants for Computing Activities, February, 1971, p. 5) These activities will of course be carried on.

SELECTION OF CRT FOR TEXT-EDITING

A good system for text-entering and text-editing meets three criteria: it is 1) economic, 2) efficient, and 3) general. By "general" I mean that its programmed interactive relationship with the rest of the system should accommodate as many operations for retrieval, analysis, and presentation of text as economy and efficiency will allow. In other words, I seek, by concentrating on what is essential, to get the most text-handling capacity for the least expense in equipment cost and central processing unit (cpu) time.

Existing text-editing systems, it is apparent, fall short of these criteria. According to van Dam and Rice ("On-line Text-Editing: a Survey," ACM Computing Surveys, September 1971, pp. 93-114), two basic kinds of systems occur today: program editors and free-form text editors. Program editors (that is, methods for editing computer programs) abound, but none are efficient text-editors because they assume that the one-line statement is the unit operated on and because they depend on statement numbering conventions in their search routines. The paragraph, not the line, is the basic unit of free-form text, and insertion or deletion of any word or phrase usually causes a change in the content of every line in the paragraph. Program editors also consume cpu time.

Van Dam and Rice discuss seven free-form text editors. The first is the stand-alone IBM Magentic Tape Selectric Typerwriter (MT/ST), which I will not consider here because it does not interact. Its disadvantages as an editor, however, are treated in "The Production of Machine-Readable Text," herewith submitted. Their second is the Astrotype system which achieves MT/ST editing efficiency by attaching typewriters to a PDP-8. Although a PDP-8 has a more versatile cpu than an MT/ST, Lawrence can do better with an editing CRT on its PDP-11. IBM'sAdministrative Terminal System for 360 computers and VIPCOM, a similar package with photocomposition output, have the same crippling deficiency as Astrotype and MT/ST: typewriter Input/output. A typewriter cannot display text instantly or show corrections as fast as they are made in the place where they are made. Space on sheets of paper is not "elastic", like that on CRT screens.

Van Dam and Rice's report on CRT-based systems shows that these too leave much to be desired. An elaborate system developed on a Rand Tablet at Carnegie-Mellon University, in which the editor marks the text with a light pen and the computer interprets the marks, is out of the question from the standpoint of cost, according to these authors. The Hypertext Editing System and its commercial sister FRESS are used for producing reports and they feature the interesting concept of "hypertext," a non-printable but computer-accessible version of a text in which one reads not serially but by jumps from one textual fragment to another, following a given path of thought rather than the order of lines on a printed page. Although Hypertext and FRESS have efficient interactive insert and delete routines, these are based in the cpu. Moreover, the hypertext feature would be an expensive overspecialized luxury from our standpoint. NLS at Stanford, the final editing system discussed, is another sort of hyper-text adventure using the idea that a text's logical outline can be expressed as a tree structure. As implemented, the workaday insert-delete features have been slighted in favor of these grander things. The creation of text or the creative process itself seem to be the purpose of these systems, rather than the editing and analysis of text.

Van Dam and Rice omit mention of commercial text-editing systems used by the printing industry in the preparation of copy for automatic typesetting, such as the famous offline system marketed by Harris-

Intertype and DEC's Typeset-8 system, which in fact is now implemented on the PDP-11. Since justified multi-font hardcopy output having the quality of printed text is not required for the production and study of machine-readable texts here proposed, these systems offer no solution. The DEC system moreover is typewriter-bound.

Modern CRT's, developed originally as keypunch replacements but now beginning to transcend the notion that all data is divided into fixed fields on 80-column records, both release text from confinement to hard copy and free the cpu from incessant editing drudgery, leaving it with only the occasional task of getting and putting chunks of raw and edited text and making room for the development of generalized routines for retrieval, analysis and presentation of the texts with which the CRT-user interacts.

In January 1971 I surveyed the CRT's of 49 manufacturers in the U.S. At that time Atlantic, Courier, Datapoint, Infoton, Imlac, Spiras, Sugarman and Sys manufactured CRT terminals having hardwired editing functions meeting most of the requirements for an efficient system as conceived of in this proposal. (See "Production of Machine-Readable Text", pp. 39-40, for a full discussion of these requirements). The field is undoubtedly more populous now. To edit efficiently, a terminal must have a full upper and lower case character set and easy-to-operate insert, delete and overwrite functions; it must be able to adjust text automatically to editorial changes as they are made; it must display an average-sized paragraph and store a substantial amount of text in a buffer of its own, scrolling it up and down the screen as needed. And of course it must refresh its own screen and communicate easily with the cpu. Which of many candidates will best serve our needs cannot be ascertained without another survey.

TEXT-EDITING SYSTEM

In the following paragraphs an attempt is made to sketch in the main outlines of the system. I hope that at least some of the ideas will prove a feasible basis for programming. It seems logical that the routines for getting text for the CRT and disposing of it after processing may be gradually expanded into information retrieval and text analysis routines. The reason is that textual analysis may be seen as a special case of information retrieval. If this special case is kept in mind, development of retrieval and analysis routines may go hand in hand.

Six commands, four defined locations and one search parameter format should give sufficient scope for the operations I have in mind. The commands are FIND, SHOW, SAVE, SHELVE, CUT, and PRINT. The locations are TEXT, SAFE, SHELF, and PRINTER: the last is necessary because there will be various ways to print output, so that a choice must be made. TEXT is the input document; SHELF is temporary storage, and SAFE is the resting place of finished work. I imagine that all three files will reside temporarily on the disc for processing, but that users will specify DEC (tape) or DISC as the apparent location or desired ultimate destination of TEXT and SAFE. I hope that finding space and remembering where it is can be the responsibility of the system.

FIND requires IN TEXT, IN SAFE, or ON SHELF to designate which file to search, and having searched, it reports the number of hits, to allow for the user to try something else before commanding SHOW. Whatever is SHOWn may be SAVEd, SHELVEd, or CUT. It there is too much data to show, the system allows the user to scroll it through the screen until it is exhausted.

FIND also requires the definition of a PATTERN to be searched for, and here it may be possible to introduce a great deal of generality. The pattern is a character string which may have a definite length, or an indefinite length; it may have gaps in it of definite or indefinite length; spaces are treated as characters. If we define the pattern graphically - that is, by simply creating its image on the screen - any user will be able to describe what he wants clearly and simply without having to use any arbitrary and difficult mathematical notation, or even having to count anything. In my illustrative notation 'x' stands for any character, the conventional '...' stands for undefined gaps in text, and quotation marks delimit the pattern. If real x's, periods, or quotation marks are needed in the pattern, a code must be set beside them to indicate that they are real. A command for finding all 43-character spans containing "day" would look like this: FIND IN TEXT PATTERN "xxxxxxxxxxxxxxxxxxxxdayxxxxxxxxxxxxxxxxxxxx". This pattern would of course catch "Friday", "daylight," and so forth. If you wanted only the word "day" the pattern would be "xxxxxxxxxxxxxxxxxxx day xxxxxxxxxxxxxxxxxxx".

By using SHELF the interacter can produce most of the effects of Boolean logic. For example, if he wanted all 23-character spans containing "night" and "day", he could subject his "day" file to this FIND pattern: "...night...". Sentences containing "night" and "day", if the text expresses sentence endings by a period plus two spaces (differentiating it from the period in an abbreviation), can be found by this pattern: ". x... night ... day .... x", assuming that the system can interpret the meaning of '. ', and '.... '. The pattern 'xxxxxxxxxxxxxxx x...x xxxxxxxxxxxxxxx' will give you every word in the text in a context of 30 characters. The result could be sorted by the 360 to form a QWIK index.

Boolean 'or' presents no problem if multiple searches are an acceptable method of securing it, and the Boolean 'not' can be achieved with the CUT function. Since the CRT has its own hardwired delete function, the software CUT will be used only for large segments of text. Whenever it is executed, the computer makes a new file on the disc adjusting the text to the deletion, and if the excision is in the TEXT tape the computer writes a new file on the tape, or if necessary asks for a new tape to put it on.

A search routine of this complexity operating serially on large amounts of free text will of course be quite inefficient. Several measures may be taken to speed up response time. 1) The user may work on one small bit of text at a time and later merge results. 2) Search patterns may be designed so as to put a rare character first, the assumption being that the computer does not try for a match until it finds the first character in a pattern. To take advantage of rigid structure in a text (as in a bibliography), the 360-44 can be programmed to interpret and pretag classes of items. Unstructured text may be tagged quite easily on the CRT. The pattern statement may use the conditions naturally delimiting sentences and paragraphs as an aid in finding these units efficiently. If sentences or paragraphs are numbered, search time will be drastically cut. 3) Preprocessing on the 360 (as in a concordance of free text or breakdown of structured text) may be used to split data into records of fixed fields. Search routines may then skip everything but the field in question, using the pattern statement to define the position in the record of the field. A line or page of free text can be treated as a fixed record. If lines or pages are numbered, enabling the computer to skip to the unit required, search time is drastically reduced.

SAVE and PRINT will require minimal programming, because no formatting will take place in this system except on the CRT screen. SAVE copies the format of the text on the screen. If more formatting is required than what can be done "manually" on the screen, it can be done on the 360, providing that automatic formatting is possible. PRINT copies the format of SAFE, so that the user will always get hard copy looking exactly like what he has seen or made on the screen, to the extent that the printer's character set can express it. If the commands can be kept simple, the user may be given a good deal of housekeeping responsibility. Gratuitous prompting can be dispensed with. The user will be protected from mistakes but he will not be notified about them until after he has made them. He will, for example, be expected to remember that he must define SAFE and won't be told about it unless he forgets. There would be less work if the computer had never asked in the first place, but waited to see. Perhaps two error messages would suffice for all situations: WHAT? when the statement is faulty or the command can't be executed, and TOO MUCH when space is lacking. A user who doesn't understand the logic of the system won't use it well, and minimal messages will help him learn. Elaborate identifying, dates, timing, diagnostic printouts, etc. can be dispensed with, too. The user need not even name his files. He must keep track of everything himself as he does in his office.

Although it is only semi-automatic, this system will perform most of the common functions of text-production and text-analysis, and combing them produce sophisticated results by a method that recognizes that no two scholars approach a text the same way and that no two texts respond equally well to the same treatment. No more need be said about the CRT as text-entering and text-editing device, but it may be noted that the system here proposed allows a writer to simulate cut and paste *composition* with much neater results, by retrieving chunks from TEXT and SHELF in the desired order and putting them in SAFE. The illustration of pattern statements for 'day' and 'night' also shows how *collocation studies* may be performed on the system.

Collation of texts may be achieved using sentences of TEXT as a pattern for FINDing sentences of the second version on SHELF. One does this by inserting FIND ON SHELF before sentences of TEXT as they scroll through the screen, being sure to put quotation marks around each sentence used as a pattern. If TEXT is double-spaced, providing for hits to be announced in the empty line, then TEXT with annotations can be deposited in STORAGE. If sentences were numbered beforehand in both texts, either manually on the CRT or by a batch program on the 360-44, collation would speed up greatly. Then, when a miss occurred, the sentence on SHELF having that number could be found and shown interlinearly for comparison.

If two different texts are assigned to TEXT and SHELF, *correlation* of their contents in many different ways would be possible, using the image of TEXT on the screen as a ready-made pattern. The system used thus could deal with the problem of *unknown authorship* by searching the unknown on SHELF for stylistic traits found in the image of TEXT. *Stylistic analysis* usually requires *syntactic analysis*. Both are facilitated, therefore, by the efficiency with which grammatical types can be identified and tagged on a CRT with hardwired editing. *Statistical analysis* is enabled by virtue of the HIT messages after each FIND command. One simply SAVEs contents of the screen to compile the statistical facts as they are called for and obtained.

The system allows for a flexible system of *information retrieval*, it goes without saying, and by judicious formation of the pattern statement and application of the device of querying the results of previous queries, very precise search parameters can be achieved. LSIB, which has programs that convert structured text into fixed-field records of classes of data and select from these fixed fields with and logic, would very much benefit from the or and not logic available on the CRT system, its nested multiple search capability, its flexible pattern statement. Using the CRT with SHELF and STORAGE, we could compile more exhaustive and precise answers to queries than machines are capable of simply because the CRT system gives free rein to human intervention in the search process. A more immediate benefit would be implementing the large mass of unstructured material in The London Stage that phase two of the project proposes to cope with. This material, consisting of parenthetical remarks and extensive "Comment" sections, can be made into a KWIK index on the 360-44 and queried by our CRT system with the utmost precision and completeness.

IMPLEMENTATION

Lawrence University owns a PDP11/20 and shares by formal agreement the 360-44 at the Institute of Paper Chemistry on the South Campus. DEC's RSTS time-sharing system, servicing 5 teletypes, is implemented on a disc operating system and two DECtape drives. At present the system provides hard copy output only by teletype. The text-editing system described above will be set up on the PDP11, whose configuration, design, and operating system are amenable to text-processing. We cannot implement it on the 360-44 because this computer uses a RAX time-sharing system which converts all lower case input to upper case. IBM is apparently not going to support development of the system to accept lower case, and we cannot expect our computer staff to undertake the task without IBM support. However, the 360-44, with its 128k memory, three disc drives, and three tape drives is necessary for large batch text-processing jobs. It can prepare text for the PDP11 text processor and information retrieval on a grand scale, as with the London Stage. Since DECtape is incompatible with the 360-44 and since actual telecommunication between the two computers has not been developed and would be expensive if it were, the best solution is to install DEC's industry-compatible tape drive on the PDP11, and transfer data from one computer to the other by tape. This configuration is assumed in the design we propose for the text-processing facility.

It is easiest to outline the steps involved in converting a volume of the London Stage into form suitable for text-editing on the PDP11/20 and for batch processing research on the IBM 360-44. These steps indicate the plans we have for computer handling of the London Stage, and they can also be generalized for other editing and research projects handling major amounts of text.

1. CONVERSION OF PRINTED TEXT TO MACHINE-READABLE FORM.

   The original text of London Stage, which consists of eleven volumes, each containing approximately 2,000,000 alphanumeric characters (including blanks, punctuation, etc.), is now being prepared for optical scanning by China Data Systems of Hong Kong and will be scanned and converted to IBM-compatible, 800 bpi, 9 track magnetic tape by International Control, Inc., Kansas City. This portion of the LSIB research has been funded by The National Endowment for the Humanities and matching private grants.

2. TRANSFER OF TEXT ON TAPE TO RAPID ACCESS DISCS.

The contents of the tapes of original text are then copied onto disc cartridges which can be read and written upon by Lawrence's PDP11/20. We plan to place approximately 1.0 million characters of original text on each disc cartridge; thus, two cartridges will contain one volume of London Stage original text. The remaining 1.4 million bytes will be available to the user; it is in this area that he will place the edited version of his 1.0 million characters of original text. (We are implicitly assuming an average increase in number of characters as a result of editing of no more than 40%, obviously.)

To accomplish this step we request funds for purchase of several items of hardware:

(a) a DEC magnetic drive (TU 10)
(b) a tape controller (TU 11)
(c) a 1.2 million word DECpack removable disc cartridge drives (RK 11)
(d) a controller for up to 8 DECpack disc cartridge drives (RK 11)
(e) two 1.2 million word disc cartridges (RK03-KA)
(f) Asynchronous Interface (KL11)
(g) a large system capability option (KH11-A)

In effect, these items complete a system of tape to disc to tape conversion. With the PDP11/20 editing from disc is far more efficient than from tape.

3. EDITING SYSTEM

The editing system requires an intelligent CRT which will facilitate access to the discs. We plan to interface an IMLAC or other suitable CRT with the PDP11/20. In addition, the system will require the design, testing, and implementation of the software required for text editing and for supervision and maintenance of the disc files of original and edited text during the editing process. We estimate four more months of programming ($4,400, staff time including benefits) over the two year period of the grant. We request funding for half this amount; Lawrence University will support half. It is our intention and present expectation that the editing and supervision software will be written in BASIC, will be fully documented, and will be made available to the educational community at a cost not to exceed the cost of reproduction of the documentation, plus listing of the programs or punched paper tape of the programs, as appropriate.

To the extent that they are not required by LSIB, the disc cartridges can and will be used by other researchers and students at Lawrence after the edited text they contain has been copied onto the tapes of edited text. All magnetic tapes and destructible disc files will be backed up with at least one archive copy. All tape labels and formats will be ANSI/industry standard, to facilitate transportability to the installations of other researchers.

RATIONALE FOR THE ABOVE APPROACH

The approach to the implementation of LSIB described above has been chosen to meet the several goals of LSIB in a cost-effective manner. Those goals are:

- The production of a data base consisting of a fully edited text of London Stage which can be readily transported to most computer installations.
- The improvement of a set of software, properly documented,

which will be useful to researchers at Lawrence and elsewhere in examining the London Stage edited text and other similarly-edited texts, with proper provision for transportability of the software to other installations.

- The production of a set of software in BASIC which will be fully use-tested in the editing of London Stage at Lawrence and will then be documented and available for dissemination to interested educational users.

It is our intention and present plan to document both the LSIB data base and the editing and research software in accordance with the standards now being established by Project CONDUIT. This effort will be coordinated by the Director of Computing Services at Lawrence, who, as Chairman of CONDUIT/EBCDIC, the CONDUIT curriculum committee in economics and business, is in regular contact with CONDUIT Central at Duke University. In order to meet these goals, we have made several choices:

- We have chosen to write our software in standard dialects of commonly-available languages (BASIC and PLI).
- We have chosen to use our conversational capability on the PDP11/20 in the text editing process, where a large amount of human intervention is both necessary and desirable, and to use our batch processing capability on the IBM 360-44 in the research process, where user requirements can be largely specified with considerable accuracy in advance of a job run, and where the requirements can be logically and/or computationally complex and still be processed efficiently from the point of view of CP time requirements. (As is well known,, complex alphanumeric searches of large data bases can place an intolerable load on a time-sharing system

the size of a PDP11/20.)

- We have chosen to obtain an industry-compatible tape drive, which will produce an important interface between the PDP11 and IBM 360-44 systems at Lawrence, enabling us to extend the benefits of both computers.
- We have chosen to develop a basically disc-oriented rather than tape-oriented text editing system because of the relative speed, reliability and, in both human and machine terms, cost-effectiveness of the disc-oriented system. We have chosen to request funds for moving-head discs rather than the more expensive fixed-head discs because the moving-head discs provide greater flexibility, less inconvenience to other users of the PDP11-20 at Lawrence, and sufficient speed for the needs of LSIB editing. Upon completion of the text editing work of LSIB, which will occur long before the end of the useful life of the moving-head disc hardware, this equipment will be available for student and faculty use on other projects, and will have educationally beneficial effects at Lawrence beyond the LSIB project itself.
- Finally, for all the reasons developed above, we have chosen to interact with the system from anintelligent CRT that will extend the capacity of the editing system and simplify the problem of terminal-computer interface.

LAWRENCE UNIVERSITY COMPUTER SERVICE STAFF

The computer staff is composed of five people. The Director Michael Hall, recently appointed, takes up his appointment August 1, 1972. He is assisted by two programmer-analysts, Walter Brown, who concentrates on academic and systems programming and Robert Nasternack, who handles administrative systems and programs. David Dreissen, an operator who is developing as a programmer and Bonita Hilgemen, a key punch operator, round out the staff. In addition a number of students work as part-time programmers. With the expansion of student programmers to handle some of the rountine programming tasks, the permanent staff will assume the added programming responsibilities entailed by this proposal.

Budget for Establishing a Text-Processing Facility at Lawrence University as Part of the *London Stage* Information Bank Project

Time: November 1, 1972 - December 31, 1974

Salaries:	Request from NSF	Lawrence Cost Sharing	Total
Programmer for CRT interface and text-processing system (4 man months @ $1000)	2000	2000	4000
Secretarial support for publicity and communication (2 man months @ $650)	650	650	1300
Benefits at 10%	265	265	530
Permanent Equipment:			
(All DEC unless noted. Cost are based on recent quotes and include installation charges)			
TU 10 DEC Magnetic Tape Drive	7350		7350
TU 11 DEC Tape Controller	3240		3240
RK 05 1.2 million word DECpack removable disc cartridge system	5360		5360
RK 11 Controller for RK 05 DEC-pack disc cartridge drives	6140		6140
RK 03-KA 1.2 million word disc cartridges: 2 @ $150	300		300
KL11 Asynchronous Interface	460		460
One CRT Terminal with some internal capabilities: IMLAC or equivalent, to be determined at time of purchase	9350		9350
Maintenance of Equipment:			
(Estimated maintenance charges for two years based on quotes or industry averages)			
TU 10 Tape Drive	850	850	1700
TU 11 Tape Controller	350	350	700

RK 05 Disc System	720	720	1440
RK 11 Disc Controller	480	480	960
KL 11 Interface	72	72	144
CRT Terminal	1000	1000	2000
Travel:			
Travel for installation of equipment	175		175
Travel for presentation of papers	250	250	500
Administrative Expense:			
Cost of reproduction or printing of papers and communi cations	300	100	400
Indirect Costs:			
@58.4% of salaries and wage	1702	1702	3404
	41014	8439	49453

Set forth below is the budget for the completion of The London Stage Information Bank, submitted to the National Endowment for the Humanities and approved by them. If this second proposal is approved by the National Science Foundation, the starred items will be affected as noted below.

Budget for Completion of
The London Stage Information Bank

Time: 1 August 1972 - 1 September 1974

Salaries:	Request	Lawrence Cost Sharing	Total
Director (1/2 time, 2 yrs.)	$17,500		17,500
Director (3 summers, 6 months)	11,700		11,700
*Programmer (1 year)	11,000		11,000
Fringe benefits	2,850		2,850
	$43,050		$43,050
Supplies: Expenses, Equipment			
Postage & Telephone	400		400
Xeroxing	200		200
Literature, consultation	300		300
*Discs, tapes, paper for computer	700		700
*Rent editing terminal (26 months)	11,000		11,000
Convert bibliographies	2,500		2,500
*Computer time (150 hrs. @ $100/hr.)		15,000	
	15,100	15,000	30,100
Travel:	1,000		1,000
Indirect costs @ 58.4% of salaries and wages*	25,000		25,000
Grand totals	$84,150	15,000	$99,150

*Authorized by NSF in letter dated 8 June 1970.

Note:
1. The budget for the programmer will be reduced by one-third since four man-months of programming are presented in this proposal.
2. Two discs at $300 will be purchased under the proposal to NSF, reducing the item for computer supplies to $400.
3. The entire budget for rent of an editing terminal ($11,000) will be deleted.
4. Lawrence will continue to supply computer time from its existing facilities on campus and at The Institute of Paper Chemistry. Some portion of the computer time should be attributed to Lawrence cost sharing in this proposal to the National Science Foundation, but the method of allocation is not clear. We prefer simply to note the Lawrence contribution here.

BIOGRAPHY OF PRINCIPAL INVESTIGATOR

Ben R. Schneider, Jr. was born in Cincinnati, Ohio on July 7, 1920, grew up in the Boston suburb of Winchester, Massachusetts, and went to college at Williams in the western part of the state, concentrating in English. After World War II, during which he was a radar man in the Army Signal Corps stationed in the Pacific, he obtained the Ph.D. at Columbia (1955) specializing in English Romantic poets. His studies for the degree included a year as research student at St. John's College, Cambridge, for the purpose of gathering material on Wordsworth's undergraduate days at the same college.

After teaching at the University of Cincinnati, the University of Colorado, and Oregon State College, he came to Lawrence University in 1955, where he now teaches eighteenth century English literature and Romantic poets. Besides articles on Wordsworth and the London Stage, he has published

Wordsworth's Cambridge Education (Cambridge University Press 1957);
The Ethos of Restoration Comedy (University of Illinois Press 1971);
and with Herbert Tjossem _Themes and Research Papers_ (Macmillan 1961).

As far back as 1964, he became interested in producing a machine-readable version of _The London Stage_. In the course of research for _The Ethos of Restoration Comedy_ he needed to know whether casting practices emphasized the character types that continually occurred in his reading of the plays. He soon found himself sorting 30,000 IBM cards, each one consisting of a performance of a role by an actor in one of the 83 plays on which he based his study. Sometime in the mid-sixties he broached the subject of a machine-readable version of _The London Stage_ to Professor G. W. Stone, Jr., Dean of the Graduate School at New York University.

Basically the problem was that this great reference work organized casting and other information in a chronological listing of performances, whereas most users of the source sought information on particular actors, plays, or roles. Getting this sort of information out of The London Stage was a bit like using a telephone book to obtain a list of the residents on a particular street. Dean Stone, and in due course the other editors of The London Stage and its publisher, Vernon Sternberg of the Southern Illinois Press, agreed that computer access to the work would considerably increase its usefulness, and in April of 1970, Dean Stone asked him to direct a project to achieve this end. His computing activities since then are fairly well covered in 2 newsletters herewith submitted, and on pp. 5-11 of this proposal.

ACCOMPANYING DOCUMENTS

1. Prospectus for completion of The London Stage Information Bank
2. Final Report on Phase One of The London Stage Project
3. "The Production of Machine-Readable text: Some of the Variables," Computers and the Humanities, September 1971
4. "The Coquette-Prude as an Actress's line in Restoration Comedy during the Time of Mrs. Oldfield," Theatre Notebook, Summer, 1968.
5. Newsletter no. 1
6. Newsletter no. 2