P. Tasman

Literary Data Processing

Abstract: A method is presented for rapid compilation of analytical indexes and concordances of printed works, using either a conventional punched-card system or an electronic data processing machine. A detailed description of the procedures used in automatically analyzing and indexing the *Summa Theologica* of St. Thomas Aquinas is given. Reference is also made to the indexing of the Dead Sea Scrolls using an IBM 705.

Introduction

Data processing generally refers to the broad concept of handling numerical data to produce accounting and statistical results. *Literary* data processing is a more recent extension of this term to include data-processing techniques adapted to the requirements of literary analysis. Mechanization of information retrieval and analysis has become an increasingly important need of scholars and researchers in the humanities who until recently have had to resort almost exclusively to manual methods of compiling analytical indexes and concordances. In fact, the careful indexing of literary works by manual methods has been a task of such magnitude that few comprehensive attempts have been made, and those have produced only limited results. Although concordances of literary and scholarly works have been manually compiled and printed over the years, thousands of others are still badly needed throughout such fields as philology, philosophy and theology.

Traditional manual methods of linguistic analysis have always been prone to inaccuracies due to human error in handling large volumes of statistical data and in the intercorrelations of these data with the individual items in context. The use of the latest data-processing tools developed primarily for science and commerce may prove a significant factor in facilitating future literary and scholarly studies.

One of the first comprehensive inquiries into an automated or mechanized method of processing literary information came about as a result of nearly four years of manual research by the Rev. Robert Busa, S. J.,* on the use of the preposition *in* throughout the writings of St. Thomas Aquinas. This rather typical linguistic problem involved copying and classifying thousands and thousands of text passages onto cards and comparing their content. In seeking ways to expand this detailed work, he decided it was necessary to seek a modern, "mechanical," way to produce the necessary analytical tools.

Literary data-processing objectives

The primary objective of the literary data-processing project was to provide a method by which any text could be machine-analyzed and indexed down to its simplest meaningful elements, the words. After reduction, the words were to be compiled in a variety of specialized lists, the number and criteria to be determined by the ultimate requirements of the specific study.

In reducing the original text to words scholars had the responsibility of properly identifying the location of each word and it relationship with associated words, phrases, paragraphs and thoughts. To accomplish this result, it was necessary to establish a set of analytical ground rules applicable to the individual research and consistent with the degree of refinement desired in analyzing the text material. This reduction and the ground rules for the St. Thomas Aquinas project will be fully described.

---

*Father Busa had begun his work in 1941 as a member of the Pontifical Faculty of Philosophy at the Aloisianum located in Gallarate near Milan, Italy. This school has been devoted to the study of the writings of St. Thomas Aquinas and associated works for more than 130 years. Father Busa discussed the project in 1949 with Thomas J. Watson, Sr., who assigned technical assistants to explore the possibilities of mechanizing this work. Soon after the project began in 1949, it received the support of Francis Cardinal Spellman, who foresaw the considerable benefit to be derived from mechanizing statistical literary research. The process and methods described here were developed by the author in collaboration with Father Busa.

Definition and uses of a concordance

A concordance is an alphabetical collection of the individual words used by an author in a given work, citing every passage in which each word appears. Literary concordances are not a final objective in themselves, but are simply useful tools for more creative scholarly work. Two typical applications of concordances are: the analysis of one language into the common roots and derivations of different languages; and comparative psychological literary research based on the frequency and structure of certain ideas and situation patterns in the works of various authors. Concordances also find use in the following fields:

• *Theological studies*

With a good comprehensive concordance, there is a practical way for theological scholars to study the meaning and value of thoughts expressed in the great religious works, for example, the meaning of the word *mercy* as used in both the Old and New Testament or the meaning of the word *grace* in its varied uses.

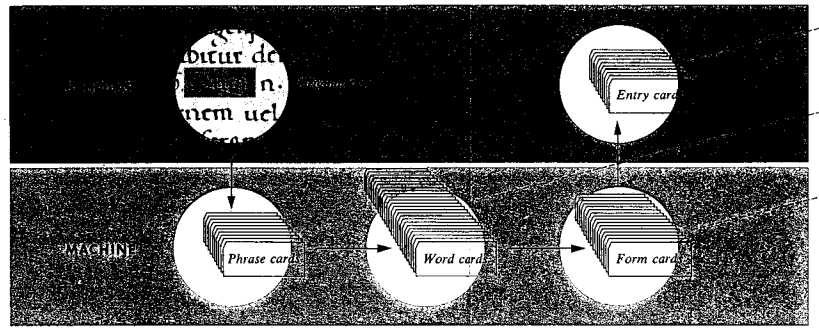
• *Philosophical studies*

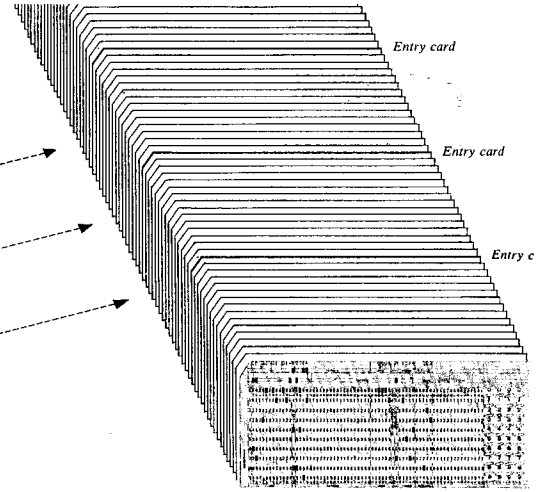
Historical research on general vocabulary is facilitated as are studies of the vocabulary of one philosopher or of several. Philosophical studies may also be used as aids in theoretical researches into the structure and the elements of the common language.

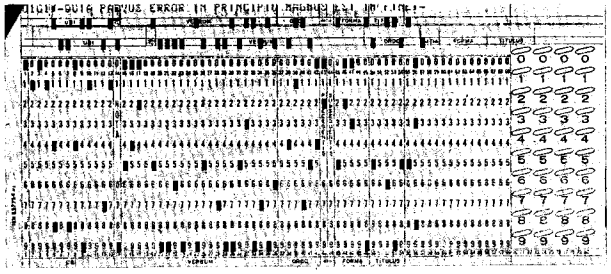
• *Philological studies*

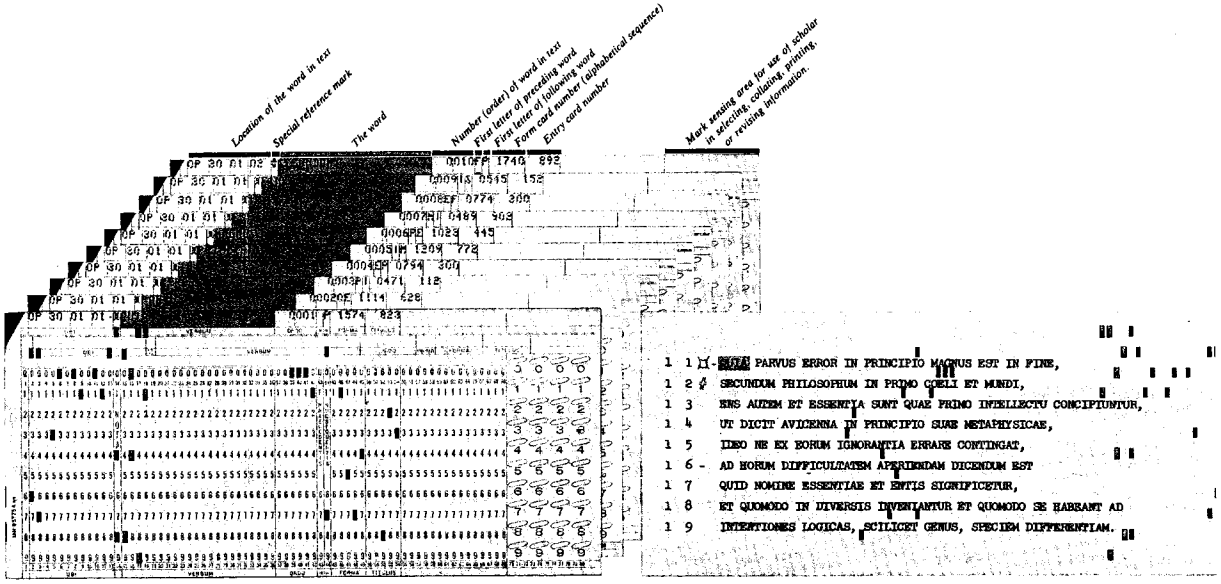
First among philological studies is lexicography, the preparation of a lexicon or dictionary.

In preparing a lexicon, a concordance provides the necessary material to study and determine the various meanings of the words which appear in the lexicon. Following this comes grammatical research on the forms of the words and the grammatical structure of the language. Other extensions of philology may involve studies comparing medical or technical words in several languages or among several disciplines.

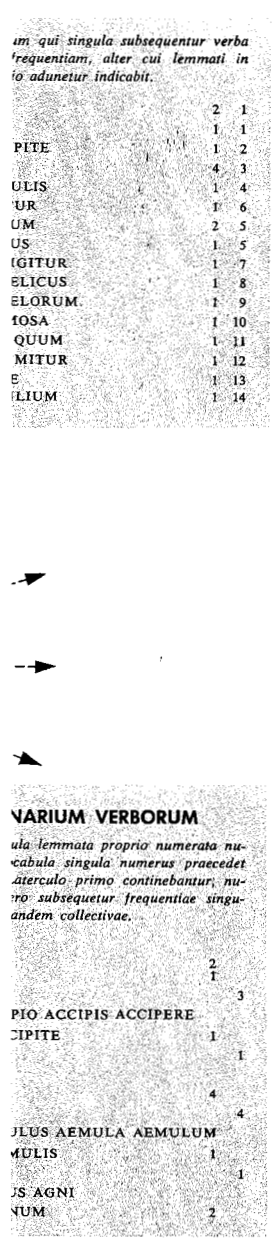








Quia parvus error in principio magnus est in fin



im qui singula subsequentur verba frequentiam, alter cui lemmati in io adunetur indicabit.

	2	1
	1	1
PITE	1	2
	4	3
ULIS	1	4
UR	1	6
UM	2	5
US	1	5
IGITUR	1	7
ELICUS	1	8
ELORUM	1	9
fOSA	1	10
QUUM	1	11
MITUR	1	12
E	1	13
LIUM	1	14

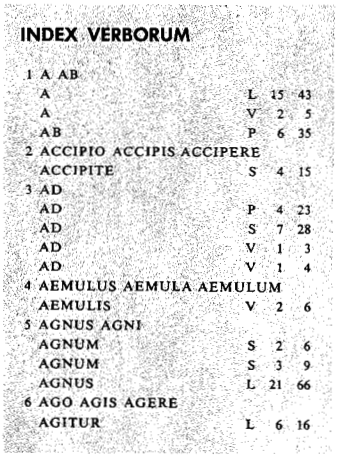
VARIUM VERBORUM

ula lemmata proprio numerata nu- cabula singula numerus praecedet aterculo primo continebantur, nu- ro subsequetur frequentiae singu- andem collectivae.

	2
	1
	3
PIO ACCIPIS ACCIPERE	
CIPITE	1
	1
	4
	4
JLUS AEMULA AEMULUM	
MULIS	1
	1
JS AGNI	
NUM	2

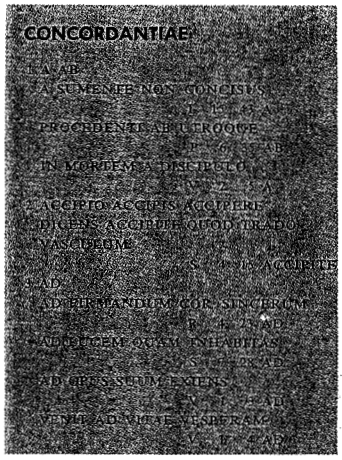
CONSPECTUS LEMMATUM RATIONARII

1 A AB
2 ACCIPIO ACCIPIS ACCIPERE
3 AD
4 AEMULUS AEMULA AEMULUM
5 AGNUS AGNI
6 AGO AGIS AGERE
7 AMBIGO AMBIGIS AMBIGERE
8 ANGELICUS ANGELICA ANGELICUM
9 ANGELUS ANGELI
10 ANIMOSUS ANIMOSA ANIMOSUM
11 ANTIQUUS ANTIQUA ANTIQUUM
12 ASSUMO ASSUMIS ASSUMERE
13 AUDEO AUDES AUDERE
14 AUXILIUM AUXILII
15 AZYMUS AZYMA AZYMUM
16 BELLUM BELLI
17 BENEDICTIO BENEDICTIONIS
18 BIBO BIBIS BIBERE



INDEX VERBORUM

1 A AB				
A	L	15	43	
A	V	2	5	
AB	P	6	35	
2 ACCIPIO ACCIPIS ACCIPERE				
ACCIPITE	S	4	15	
3 AD				
AD	P	4	23	
AD	S	7	28	
AD	V	1	3	
AD	V	1	4	
4 AEMULUS AEMULA AEMULUM				
AEMULIS	V	2	6	
5 AGNUS AGNI				
AGNUM	S	2	6	
AGNUM	S	3	9	
AGNUS	L	21	66	
6 AGO AGIS AGERE				
AGITUR	L	6	16	



CONCORDANTIAE

Varia Specimina Concordantiarum
Sancti Thomae Aquinatis Hymnorum Ritualium
—Roberto Busa

General considerations in compiling a concordance

The following five stages are considered most significant in the compiling of a concordance using automated techniques:

1. Analysis of the text into *thoughts* (logical paragraphs).
2. Analysis and transcription of the text into *phrases* (meaningful sub-grouping of words) of sizes suitable for machine processing.
3. Reduction of these phrases into single words, *all* words of the text to be included.
4. Indication of the reference, placement and value of the individual words.
5. *Family* classification, alphabetizing, and indexing of the individual words. These classifications are logical semantic headings, such as a verb infinitive.
6. The physical association of the individual words with the text in all places where they appear, prepared in such form that these associations may be useful to researchers in scholarly and statistical studies.

Linguistic analysis

This is an analytical determination of the several categories under which the elements of an expressed sequence of human thought may be grouped, classified and described. Each element will be listed under its proper category. In linguistic analysis there are two major areas of study:

1. Indexes and concordances of words.
2. Literary statistics:
   a) Phonemes and letters, groups of letters and symbols; morphemes, prefixes, suffixes, endings and roots; tones and accents.
   b) Words according to their graphic and phonetic structure only, or according to their graphic or phonetic structure as well as their semantic content.
   c) Sentences, paragraph discourses and their meaningful parts.

Linguistic analysis dealing with symbolic and algebraic logic can also be considered in mechanization of literary data processing.

Mechanization in linguistic analysis

Literary "mechanization" will greatly accelerate the collecting, grouping, comparing, correlating and counting of the elements of language in any large-scale analytical project. It will also ensure far greater accuracy and flexibility. While mechanized collections of statistical analyses may easily be checked and verified by other scholars, it is not so with manually documented collections. Through mechanization, the text material also becomes rapidly explorable by machine. Word elements of the text may be handled mechanically in such a way that they are alphabetized from left to right as well as right to left, an important factor in working with inflected languages such as Latin or in producing a rhyming dictionary.

The following processes exemplify the methods recently developed and used for preparing complete card-file indexes and several special concordances of the *Summa Theologica* of St. Thomas Aquinas. It will be seen, however, through a study of the simplified chart of operations that this method is equally applicable to other areas of literary research and documentation, taking into account certain additions or modifications. Conventional punched-card machines have been utilized in the method to be described. This has been done on the premise that the system should be applicable to universities, colleges and research organizations where limited funds might be available and utilization of large-scale data-processing equipment would be impractical. We will, however, discuss the processes which have been modified to accommodate those cases where faster handling and more detail is required, using large-scale data-processing equipment like the IBM 705.

Specific phases of automation in the literary analysis of the *Summa Theologica* of St. Thomas Aquinas

1. The scholar analyzes the text, marking it with precise instructions for card punching.
2. A clerk copies the text using a special typewriter which operates a card punch. This typewriter has a keyboard similar to that of a conventional typewriter and produces the *phrase cards*.

These cards contain all the lines or phrases of the text, one on each card, in sequence, transcribed in symbols (punched holes) that can be understood by the machine. Each phrase is preceded by the reference to the place where this line is found and provided with a serial number and a special reference sign.

A second clerk types each phrase of the text on the appropriate phrase card which has already been punched, using a checking machine.

In this way the accuracy of the text cards is rigorously checked and cards containing transcription errors are replaced.

3. From the phrase cards the machine automatically produces the *word cards* and a complete copy of the text, phrase by phrase.

Each of these word cards contains only a single word of the text. This means that the machine produces for each phrase card several new cards corresponding to the single words contained in the phrase. We have now a word card file with as many cards as words in the text.

In the course of this same operation each word is also accompanied automatically by identifying data.

These data are at present the following (they could, of course, be different and more numerous):

a) The quotation of the place where the word is found. This quotation is punched in the zone indicated by UBI (where) in the sample card.

b) The first letter of the preceding word and the first letter of the following word. See in the card the item INITIA (beginning).

c) The number indicating the position which this word holds in the order of the text; for example, that it is the 18th or 121st word, and so forth. In the card see the zone ORDO.

d) The special reference mark which characterizes the phrase to which the word belongs. In the card see the zone NOTA.

The number of these special marks can be indefinite. For the *index thomisticus* four signs were adopted:

/ means: Here St. Thomas refers to other passages of his own work.

means: These are the words of another author whom St. Thomas quotes here literally.

□ means: The phrases cannot be entirely thomistic since St. Thomas quotes here according to the doctrine of others, or comments on them or refutes them, etc.

' means: Here St. Thomas refers to other passages of his own work.

The context of the same word is printed on the reverse side of each word card — in the horizontal spaces between the punched lines — with a maximum of 12 lines or 64 strokes or an average of 100-120 words.

4. From the word cards the machine automatically produces *form cards*, without the necessity of any human intervention. All word duplicates are eliminated here, leaving only one card for each *graphically* different word.

In this phase of operation a "word" is considered as a continuous sequence of symbols, delimited on both sides by a space. The words are therefore, understood here according only to their graphic structure. Thus, for example, *were, be, am,* would appear here as three different words: the first among the words starting with *w*, the second among those starting with *b* and the third among those starting with *a*. The word *lead* would be considered as a single word though it can be two words: *lead* (conduct) or *lead* (metal), as the machine does not distinguish between homographs.

Likewise the word *weren't* would be considered as a single word while *have been* would be broken up by the machine into two words.

On each of these form cards the word is automatically accompanied by the number of times in which it appears in the entire work, that is by the indication of its fre-

quency and by a number which indicates what position it occupies in the direct alphabetical sequence of all words. This number, which represents a first numerical codification of the word, will be entered automatically in the appropriate word cards at the point indicated by the item FORMA.

This result is obtained with the following steps:

a) The machine puts all word cards in alphabetical order;

b) The machine prints on sheets of paper the different words which it finds by examining all the word cards at a rate of about 6000 per hour. It prints the first word. If the following word is different from it, it prints it. If it is identical, it does not print it, but counts it. After it has finished counting all identical words, it prints the total number next to the printed word. It proceeds then to print the following different word and so forth.

While in this way it compiles the first summary and practical statistical list of the author's vocabulary, the machine punches at the same time another set of cards, one card for each *different* word. Each of these cards has its frequency and a number indicating the position it holds in the alphabetical sequence also punched into it.

The scholar examines the list of the form cards and groups all the different forms of one and the same *graphic-semantic* unit under the single expression or word which will serve as an entry listing. For example, it will assemble the words *were, are, be* under the entry indication of the verb *to be*.

In addition, the scholar separates all the homographs in their various uses. For example, in Latin *cane* can be a form of the verb *canere,* the ablative of *canis-is,* and the vocative of *canus-i.* In Italian, for example, the form *porta* can mean *door,* or *he carries* or, *take* this. (Porta qui questo.)

The scholar in this way compiles a list of the entry words for the complete word index.

A card is then punched for each of these, and these cards are arranged alphabetically and numbered sequentially.

In this way we obtain the complete file of *entry cards.*

These form a second summary list of the author's lexicon, based on the structure of the words he uses, but on a basis which is no longer purely graphic, but graphic-*semantic.*

The form cards and the word cards are then grouped under the respective entry cards. Then the second code number is automatically punched in all word cards (in the zone indicated in the sample card with the word TITULUS and in all form cards.

At the same time the machine adds to the entry card an indication of its total frequency, that is, the total number of times in which it figures in the text under one form or another.

5. If it has not been done before, all these four groups of cards can now be "interpreted;" that is the machine will print on the top of each card in letters and consecutive numbers whatever information the card contains "written" in holes.

Starting from the punching and initial checking, the machine has thus produced two complete transcriptions of the text (the phrase cards and the word cards) and two summary indexes of the terminology used by the author (the form cards and the entry cards). With this arrangement, the text is now ready for further investigations to be undertaken with the help of the machine.

The machine greatly reduces the amount of time which the scholar must use for mere mechanical research, filing, classifying and sorting. *It permits him to use more of his specific scholarly abilities, such as interpreting and organizing the information obtained.*

6. Information on any desired groups of these cards can now be printed on sheets, in brochures or books or on other cards. Valuable tools in philological research like an *index verborum* or *concordance* are available without further scholarly effort.

It is sufficient to feed into the machine the word cards inserted between the entry cards or form cards or both. It will then print one after the other in definite form, already paginated and with the suitable spacings. This may include all of the pages comprising the large volumes of the concordance.

The printing speed ranges from 4800 to 60,000 lines per hour, depending on the type of machine used.

Other features and time considerations

In many literary data-processing projects the so-called mark-sensing device can be used. The right end of this card contains many small ovoidal outlines. In these the scholar can enter pencil marks according to precodified retrieval combinations. The machine will sense the presence of the graphite on the card and punch certain corresponding holes, by which it can automatically effect all the operations of computation, selection, correlation and printing which the scholar has indicated.

The *manual* indexing of the complete writings of St. Thomas Aquinas (which contain approximately 13 million words) would take 50 scholars 40 years.

The procedures employing smaller machines will produce the indexes and concordances in a considerably more accurate manner in about 4 years with 10 scholars.

With the planned use of large-scale data processing machines, the total time required will be reduced to approximately 25% of that required using the punched-card techniques or *less than one year.* It may be stated that, in principle, and depending on how detailed a result is desired, the mechanized procedures for indexing cut the time required for the work to approximately 1/40th that of manual and duplicating processes.

The considerations for utilizing larger machines take into account the fact that reducing the text into units (single words), each one accompanied by its indicative data, and then alphabetizing all of these units is a long and bulky operation. For example, the 1,600,000 cards which have now been produced at the Literary Data-Processing Center in Gallarate, Italy, represent all the words of the *Summa Theologica* of St. Thomas Aquinas. The alphabetizing of these cards alone, considering that

they would each have to pass an average of 20 times through a sorter, would be equivalent to 32 million cards for machine processing, and while it is possible to sort these cards at a maximum speed of 60,000 cards per hour, it may readily be seen that this would be a large undertaking. It must also be recognized that but few searches have to go from one line or word or letter of a text to another word or letter without interruption. Consequently, the text must be explored in a direction which is horizontal to the normal vertical feeding of cards. This indicates the need that text for exploration be placed on continuous paper or high-speed magnetic tape.

Comparative methods analysis

Using the text material of approximately 2,000 pages from the *Summa Theologica* which has resulted in nearly 1,600,000 individual word cards, we can arrive at the following comparisons:

- *Manual method*

In this application, we have estimated three persons would be required for 20,000 hours to produce the lexicon file index and concordance covered in this paper.

- *Conventional punched-card method*

Utilizing standard punched-card equipment, three persons would be required for a total of 1,000 hours.

- *Large-scale data-processing method*

Through the use of this equipment, it has been estimated that one person working for 60 hours could produce the same results. This, of course, is exclusive of the preparation and programming time. It may readily be seen that the method, as well as the standard use of punched card equipment considerably lessen the burden now placed on scholars in compiling, alphabetizing and collating text.

Programming and indexing of the Dead Sea Scrolls

The complex programming necessary to handle the indexing of the Dead Sea Scrolls is covered in an addendum which will be available on request.

In this application, the major objective is compactness of magnetic tape files and the speed at which tapes can be read, written and printed.

Each word card which has been prepared in accordance with the preceding text explanations is initially converted to magnetic tape in blocks of records 80 characters long. The process is broken down into four runs for the IBM 705. Run 1 deals with sorting and inverting each word in memory. Run 2 deals with creating the frequency count of the sorted words and their summarization. Run 3 deals with a merging and grouping of the different word tapes with the entry word tape. Run 4 provides for collating each word tape with the original phrase tape. Each word is grouped with its relation to the particular sentence of the text where it appears.

The cards to be used in the indexing of this work are punched using the ancient Hebrew characters and will be prepared at the Literary Data Processing Center in Gallarate, Italy. To accommodate the text material, normally written from right to left, a modified card punch has been produced to permit cards to be punched in this manner rather than in the conventional left-to-right pattern. This arrangement permits the card-punch operators to punch the text exactly as they read it and the sequence to be preserved accordingly.

The Dead Sea Scrolls published to date contain approximately 50,000 words. The punching and verifying required approximately 3 to 4 weeks and the IBM 705 programming about 3 weeks. The complete description describing this work is covered in the addendum.

Conclusions

In the preceding description a detailed explanation has been given of the methods and procedures developed and found to be adaptable to literary data processing.

While the methods evolved were mainly intended to satisfy the requirements of the initial project, it soon became apparent that the work had far-reaching effects in many areas similar to literary analysis.* As an example, the application of the principles to the indexing of the Dead Sea Scrolls had to take into account the many plates in which there were incomplete words or words totally obliterated. This presented no particular problem inasmuch as the program provided for an analysis of the text, taking into account the frequency, use and sequence of words in a particular text and their context. It becomes apparent that this forms a rudimentary system for analyzing the writing style of an author and also a tool to interpolate missing words or to detect foreign additions which are uncommon to the author. While it will not be absolutely certain that exact substitution will be made, a more accurate machine substitution will be possible than is conceivable by manual methods.†

Apart from literary analysis, it appears that other areas of documentation such as legal, chemical, medical, scientific, and engineering information are now susceptible to the methods evolved. It is evident, of course, that the transcription of the documents in these other fields necessitates special sets of ground rules and codes in order to provide for information retrieval, and the results will depend entirely upon the degree and refinement of coding and the variety of cross referencing desired.

The indexing and coding techniques developed by this method offer a comparatively fast method of literature searching, and it appears that the machine-searching application may initiate a new era of *language engineering*. It should certainly lead to improved and more sophisticated techniques for use in libraries, chemical documentation, and abstract preparation, as well as in literary analysis.

*Less known and more specialized studies, referred to as literary statistics or statistical linguistics were pioneered in Italy by Prof. Marcello Boldrini of the University of Rome. Many of these investigations can be undertaken on the above mentioned cards, for example, the statistical computation of the phonemes and of the length of the words.

†Up to *five* consecutive words have been "re-written" by the data processing machine in experimental tests where the words were intentionally left out of the text and blank spots indicated.