DOCUMENT RESUME

ED 071 686 LI 004 082

AUTHOR	Mathis, Betty Ann
TITLE	Techniques for the Evaluation and Improvement of Computer-Produced Abstracts.
INSTITUTION	Ohio State Univ., Columbus. Computer and Information Science Research Center.
SPONS AGENCY	National Science Foundation, Washington, D.C.
REPORT NO	OSU-CISRC-TR-15
PUB DATE	Dec 72
NOTE	275p.; (202 References)
EDRS PRICE	MF-$0.65 HC-$9.87
DESCRIPTORS	*Abstracting; Abstracts; Algorithms; *Automation; Computers; *Electronic Data Processing; Evaluation; Periodicals
IDENTIFIERS	*Automatic Abstracting

ABSTRACT

An automatic abstracting system, named ADAM, implemented on the IBM 370, receives journal articles as input and produces abstracts as output. An algorithm has been developed which considers every sentence in the input text and rejects sentences which are not suitable for inclusion in the abstract. All sentences which are not rejected are included in the set of sentences which are candidates for inclusion in the abstract. The quality of the abstracts can be evaluated by means of a two-step evaluation procedure. The first step determines the conformity of the abstracts to the defined criteria for an acceptable abstract. The second step provides an objective evaluation criterion for abstract quality based on a comparison of the abstract with its parent document. Based on the results of this evaluation, several techniques have been developed to improve the quality of the abstracts. These procedures modify the form, arrangement, and content of the sentences selected for the abstract. The revision, deletion, or creation of sentences is performed according to a number of generalized rules which are based on the structural characteristics of the sentences. This modification produces abstracts in which the flow of ideas is improved and which represent a more nearly coherent whole. (Author/SJ)

TECHNICAL REPORT SERIES

COMPUTER & INFORMATION SCIENCE RESEARCH CENTER

THE OHIO STATE UNIVERSITY COLUMBUS, OHIO

(OSU-CISRC-TR-72-15)

TECHNIQUES FOR THE EVALUATION AND IMPROVEMENT OF COMPUTER-PRODUCED ABSTRACTS

by

Betty Ann Mathis

Work performed under

Grant No. 534.1, National Science Foundation

Computer and Information Science Research Center
The Ohio State University
Columbus, Ohio 43210

December 1972

ABSTRACT

An authomatic abstracting system, named ADAM, has been implemented on the IBM 370. ADAM receives journal articles as input and produces abstracts as output. An algorithm has been developed which considers every sentence in the input text and rejects sentences which are not suitable for inclusion in the abstract. All sentences which are not rejected are included in the set of sentences which are candidates for inclusion in the abstract.

The quality of the abstracts can be evaluated by means of a two-step evaluation procedure. The first step of the procedure determines the conformity of the abstracts to the defined criteria for an acceptable abstract for the given system. The second step provides an objective evaluation criterion for abstract quality based on a comparison of the abstract with its parent document. The second step is based on the assumption that an abstract should present the maximum amount of data from the parent document in the minimum amount of length.

Based on the results of this evaluation, several techniques have been developed to improve the quality of the abstracts. These procedures modify the form, arrangement, and content of the sentences selected for the abstract. The revision, deletion, or creation of sentences is performed according to a number of generalized rules which are based on the structural characteristics of the sentences. This modification produces abstracts in which the flow of ideas is improved and which represent a more nearly coherent whole. The abstracts also show improvement according to the objective evaluation criteria.

PREFACE

This work was done in partial fulfillment of the requirements for a doctor of philosophy degree in Computer and Information Science from The Ohio State University. It was supported in part by Grant No. GN 534.1 from the Office of Science Information Service, National Science Foundation, to the Computer and Information Science Research Center of The Ohio State University

The Computer and Information Science Research Center of The Ohio State University is an interdisciplinary research organization which consists of the staff, graduate students, and faculty of many University departments and laboratories. This report is based on research accomplished in cooperation with the Department of Computer and Information Science.

The research was administered and monitored by The Ohio State University Research Foundation.

ACKNOWLEDGEMENTS

I am indebted to my adviser, Dr. James E. Rush, for his help and advice during the preparation of this dissertation and throughout my graduate career. I am also grateful to the other two members of my reading committee, Dr. Marshall C. Yovits and Dr. Anthony E. Petrarca for their contributions to this work. My graduate studies have been enriched by courses taught by Professors Ernst, Foulk, Lazorick, Liu, Montgomery, Petrarca, Rustagi, Reeker, Rush, Saltzer, White and Yovits.

Richard Salvador and Antonio Zamora worked with Dr. Rush on the original design of ADAM and I wish to thank them for allowing me to continue work on their system. I also wish to thank Carol E. Young for allowing me to use her grammatical class assignment programs in my research. I am grateful to Dr. Bertrand C. Landry for reading one of the early drafts of this dissertation and for providing valuable suggestions. Michael McCabe helped me with some of the programming of ADAM and I am appreciative of his assistance. I also wish to thank the students in the undergraduate information storage and retrieval courses who tested the abstracting system. Thanks is also due to my typists, Mrs. Mary Kimball and Mrs. Constance Fitzpatrick.

During my graduate studies, I received financial aid from a University Fellowship and research support from a grant (GN-534.1) from the National Science Foundation to the Computer and Information Science Research Center. Computer time for this research was provided by the grant and The Ohio State University Instruction and Research Computer Center.

A special word of thanks is due my parents, my brother and Mike for their help and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

QUOTATION REFERENCE

CHAPTER I. INTRODUCTION

1. The Need for Abstracting

> One of the diseases of this age is the multiplicity of books; they doth so overcharge the world that it is not able to digest the abundance of idle matter that is every day hatched and brought forth into the world.
>
> Barnaby Rich, 1600

The disease of 1600 caused by a multiplicity of books has turned into an epidemic in the 20th century casued by a multiplicity of books, journals, monographs, newspapers, reprints, preprints, and xerox copies. The world today is inundated with a barrage of printed messages representing the ever increasing amounts of scholarly, and not-so-scholarly, research. We live in an era of rapid discovery and far-reaching exploration and we have come to expect that the results of these endeavors will be recorded and published for future use. The problem that arises from the abundance of publications stems from our efforts to digest and organize the published ideas.

The ability of individual human beings to know and understand all kinds of data has not kept pace with the amount of data that can be known. The volume of recorded data has increased because it results from the collective achievements of millions of individuals, but the mental capacity of any one individual has remained almost constant throughout the years. Any one individual can know only a fraction of the total available data; therefore, his knowledge of some data must be limited to a few key ideas. These key ideas represent an abstraction

of the total data available. There is thus an inherent need for abstraction. Without it the collective knowledge of mankind would remain virtually constant at a very primative level. But it should be noted that although abstracting itself is a necessity, the form of the abstract may vary.

2. The Need for Abstracts in the Scientific Community

In the scientific community, the problem created by the increase in the amount of available data seems especially acute. No one researcher is able to keep up with the vast amount of published research results. He must specialize and limit the number of areas of research about which he wishes to be well informed. He must choose only a few articles that he wishes to study thoroughly. Abstracts can help the scientist identify those publications which have the greatest potential value and therefore warrant his attention. The purpose of abstracts in technical literature is to facilitate quick and accurate identification of the topics of published papers. Their objective is to save a prospective reader time and effort in finding useful data in a given article or report (1). Abstracts can play a vital role both in coping with the influx of new data and in searching for older data (2).

While abstracts have proved to be valuable aids to the researcher, the tremendous increase in the amount of published research which they are supposed to ameliorate has caused problems in their production. Most abstracting services rely on subject experts to read the documents and to write original abstracts. This method of abstract production

becomes more cumbersome with each increase in the volume of documents to be abstracted. Such increases demand more experienced abstractors, more facilities and more efficient methods of handling all phases of abstract production. It becomes more difficult to maintain a uniform level of quality and exhaustive coverage of any given field. These problems are usually most clearly reflected in periodic rises in the cost of a supscription to the abstract journal. One of the appealing possible solutions to these problems is the development of computer systems to abstract documents.

3. The Production of Abstracts by Computer

Perhaps the greatest obstacle to the complete automation of the process of abstracting is our lack of understanding of how human beings are able to abstract documents. We have very little knowledge of the method of selection of certain ideas to be included in the abstract and of the linguistic processes used in expressing those ideas in a coherent abstract. There must be an increased understanding of the methods of manual production of abstracts in order to improve the quality of computer produced abstracts. As Louise Schultz states, "To assign to a machine the tasks of processing language requires enough knowledge of language to permit design of the processes and evaluation of the processing." (3) Greater understanding of linguistics will aid in the development of computer produced abstracts and the development of computer produced abstracts will contribute to a greater understanding of linguistics. The study and formalization of the intellectual task

of reading a document and writing an abstract will provide some specific algorithms that may be useful in modeling the way human beings perform the same tasks.

The production of abstracts by computer is also of interest because it represents the application of many computer and information science techniques to the solution of a specific problem. For example, the input of the text and dictionary constitutes a problem of large-file handling and storage allocation. The matching of the dictionary with the text constitutes a problem of dictionary look-up. The expression of the abstracting algorithm as a Turing machine constitutes a problem of formal language theory. The development of the abstracting algorithm constitutes a problem of semantic and syntactic analysis. The evaluation of the abstracts produced constitutes a problem of information theory.

4. The Scope of the Dissertation

"The ultimate goal," according to Wyllys, "of research in automatic abstracting is to enable a computer program to 'read' a document and to 'write' an abstract in conventional prose style, but the path to this goal is full of yet unconquered obstacles." (4) The goal of this dissertation is to gain a greater understanding of the processes needed to enable a computer program to 'read' a document and to 'write' an abstract. This study is designed to reflect an interest in the production of high quality abstracts and in the development of a workable computer-based production system.

The results of my research are reported in the four succeeding chapters. Chapter II provides background for the development of abstracting services and a review of previous research in automatic abstracting. Chapter III consists of a description of the existing abstracting programs. Chapter IV describes an objective criterion for the evaluation of the quality of computer produced abstracts. Chapter V presents methods for improving the abstracting system, including revision and creation of sentences to be included in the abstract, as well as directions for future research.

References

1. H. P. Luhn, "The Automatic Creation of Literature Abstracts", IBM Journal of Research and Development 2(2), 159 (1958).

2. R. E. Maizell, J. F. Smith and T. E. R. Singer, Abstracting Scientific and Technical Literature--An Introductory Guide and Text for Scientists, Abstractors, and Management, Wiley-Intersience, New York, New York, 1971, p. 1.

3. L. Schultz, "Language and the Computer", in H. Borko (ed.), Automated Language Processing, John Wiley and Sons, Inc., New York, New York, 1968, p. 12.

4. R. E. Wyllys, "Extracting and Abstracting by Computer", in H. Borko (ed.), Automated Language Processing, John Wiley and Sons, Inc., New York, New York, 1968, p. 128.

CHAPTER II. BACKGROUND AND REVIEW OF PREVIOUS WORK[1]

1. The Nature and Definition of Abstracting

1.1 Introduction

A discussion of the concepts "abstract" and "abstracting" is fraught with difficulty because the data upon which one must draw cover such a great span of time and because the concepts have until very recently been part of a venerable, empirical science. Thus, while the term "abstract" implies some kind of reduced form of a corpus of things, the nature of the reduction is poorly defined. And while the term "abstracting" implies a method of reduction, the known methods are all purely descriptive. Let us consider briefly the nature and significance of the traditional use of these terms.

The idea of abstraction pervades the whole of science, and the term suggests a transition from specific to general, from individual observations to class description. Thus molecular structural representation of chemical substances is an abstraction based on the detailed examination of many, but by no means all, individual chemical species. A description of some piece of research is an abstraction of the actual research. We might say therefore that an abstract is the quintessence of that from which it derives; that an abstract is a document shorn of detail. This notion is most often expressed by saying that an abstract

---

[1] The material in this Chapter will appear in the Encyclopedia of Computer Science and Technology (1) under the joint authorship of my adviser and myself.

is a condensed or abbreviated version of a document.[2]

When the term "abstract" is used alone it seems ambiguous. It may refer to the notion of an abstract property or expression ("World War II" is a concrete expression, while the word "war" is abstract). It may refer to title to real property (in this sense an abstract is a written, connected, chronological summary of the essential portions of all recorded documents and facts which can be discovered by a search of the public records of the jurisdiction within which the realty is located (2)). Or it may refer to a summary of the principal findings of the work reported in a paper (3). We are concerned here with the last of these three senses of "abstract," but it should be obvious that the three senses are really quite similar in intent, the unity of purpose being the representation of the essential qualities of the thing(s) abstracted.

1.2. Historical Development of Abstracting

In tracing the development of information storage and retrieval, it is important to ask: What is the purpose of creating records of man's experience, of abstracting, indexing, consolidating and distilling the contents of these records, of providing repositories for their housing and for central public access thereto? As Libbey (4) puts it,

---

[2] This is a very loose description and does not seem to be entirely compatible with the notion expressed in the preceding sentence. This point will be dealt with later when the term "abstract" is given a more formal definition.

> Very little advance in culture could be made, even by the greatest man of genius, if he were dependant for what knowledge he might acquire upon his own personal observations. Indeed, it might be said that exceptional mental ability involves a power to absorb the ideas of others, and even that the most original people are those who are able to borrow most freely.

Without the availability, to present generations, of records of the experiences of those who came before, one would have no foundation upon which to build; one would have always to begin afresh. This sharing of experiences is at the heart of the need for and interest in personal communication of any sort. In fact, we can say that communication is the sharing of experiences. But most of us will never have the opportunity for personal communication with many of those individuals whose experiences we would most like to share (even if we knew of the existence of such individuals). But records of the experiences of these individuals are, fortunately, made available to us and we can thus "commune" with persons long since dead or otherwise separated from us. This is the fundamental reason for the existence of libraries, of whatever form.

Concerted effort to bring scientific inquiry to an organized state seems to have had its genesis in the 17th century. According to some, Bacon (5) was most instrumental in bringing about this shaping and directing of human activity. As a result of Bacon's efforts to cause the establishment of research as a means of regenerating learning, and to found a college of research for the purpose of fostering the New Philosophy (the scientific method) and of providing for the publication of such discoveries which the research revealed, the Royal Society

(London) was eventually founded, and subsequently the whole fabric of European scientific societies was established. The Royal Society may be credited with initiating the trend for scientists to publish the results of their research. The journals where individuals first transmit these results have come to be known as the primary literature.

The Royal Society was founded in 1662 and proceedings of its meetings were first published in 1665[3]. But it was not until the 19th century, almost 200 years after the first serial publication appeared, that the collection of recorded data had grown to proportions which made it desirable to collect and summarize these data on a regular basis. In 1821, Berzelius began his Jahresberichte über die Fortschritte der physischen Wissenschaften. But Berzelius' effort was preceeded by a pharmaceutical "yearbook" which had been strated in 1795[4] with a similar purpose. However these publications were yearbooks in the sense of being annual summaries rather than more frequently appearing periodicals. But as the number of articles increased, the need for more frequently published summaries or abstracts was felt. Pharmaceutisches Centralblatt was begun in 1830 to satisfy this need and, as Figure 2.1 demonstrates, the trend toward more, more frequent, and more specialized abstracting publications has resulted in the concurrent appearance of a large number of them which presumably serve to make available the record of some portion of man's experience to

---

[3] The Philosophical Transactions.

[4] Berlinisches Jahrbuch. für die Pharmacie und die damit verbunden Wissenschaften (discontinued in 1840).

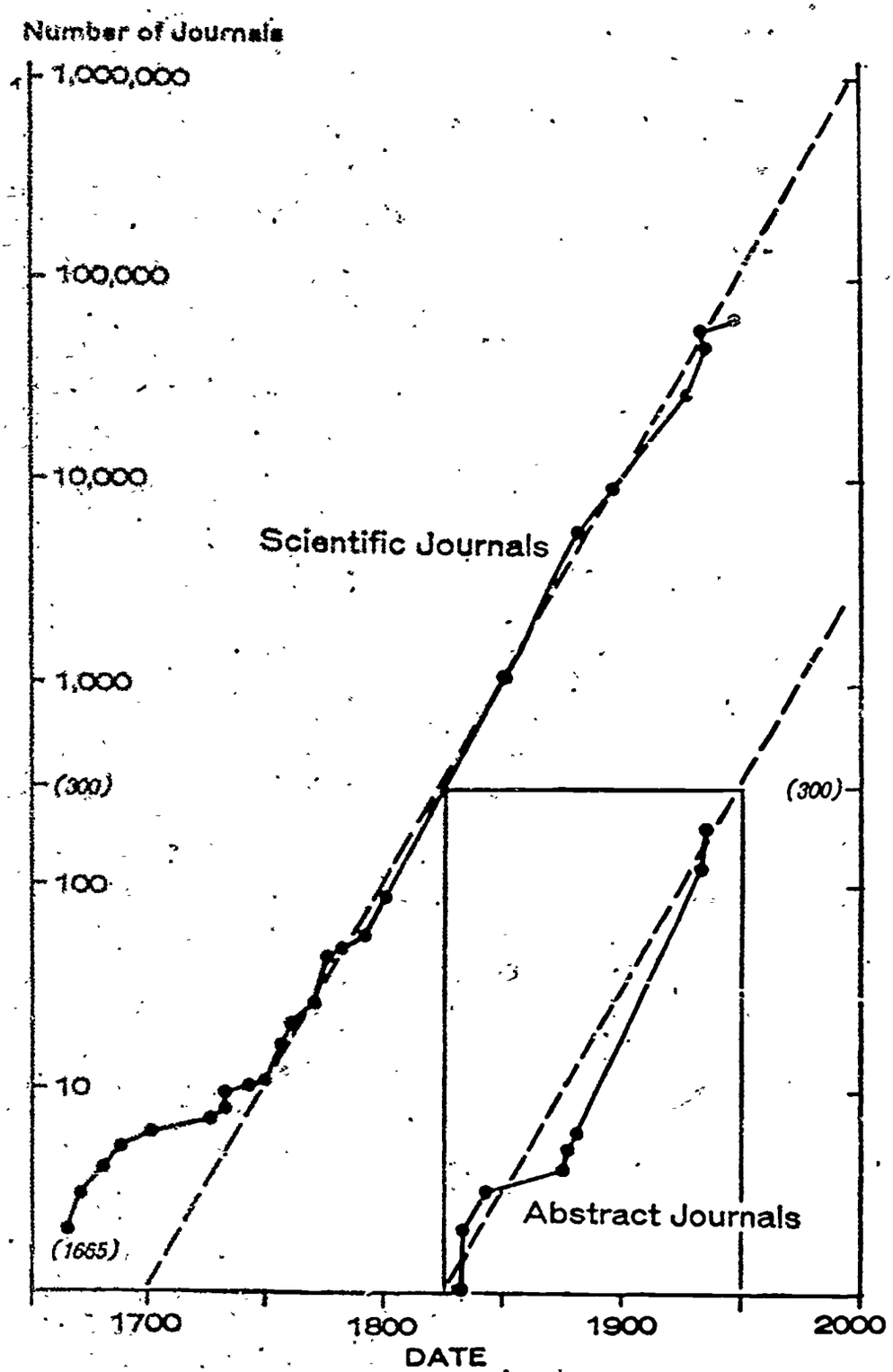


Figure 2.1 Growth in primary and secondary literature. (Reproduced from Science Since Babylon (6).)

modern scientists.[5] In fact it is fashionable to illustrate the parallel growth of the primary journal literature and the secondary (abstract) journal literature by means of a graph such as that of Figure 2.1. The parallel is striking, but the Figure raises the more interesting question: When the number of abstract journals becomes too large, what new representation of man's experience will assert itself, say around the year 1960? In other words, is there now a third level of experiential representation that is rising in importance and that will one day appear as a third "curve" on a graph such as that of Figure 2.2?

1.3. Intensional Properties of Abstracts

A handbook for authors (7) admonishes an author to be aware of the "importance taken on by his abstract."

> The ideal abstract will state briefly the problem, or purpose of the research when that information is not adequately contained in the title, indicate the theorethical or experimental plan used, accurately summarize the principal findings, and point out major conclusions. The author should keep in mind the purpose of the abstract, which is to allow the reader to determine what kind of information is in a given paper and to point out key features for use in indexing and eventual retrieval. It is never intended that the abstract substitute for the original article, but it must contain sufficient information to allow a reader to ascertain his interest. The abstract should provide adequate data for the generation of index entries concerning both the kind of information present and the key compounds reported.

Chemical Abstracts Service's Directions for Abstractors (8) states that

---

[5] I realize that one must interpret this term very broadly if all those, whom I might better call students, are to be included in the ranks of those who have need of and use recorded knowledge.

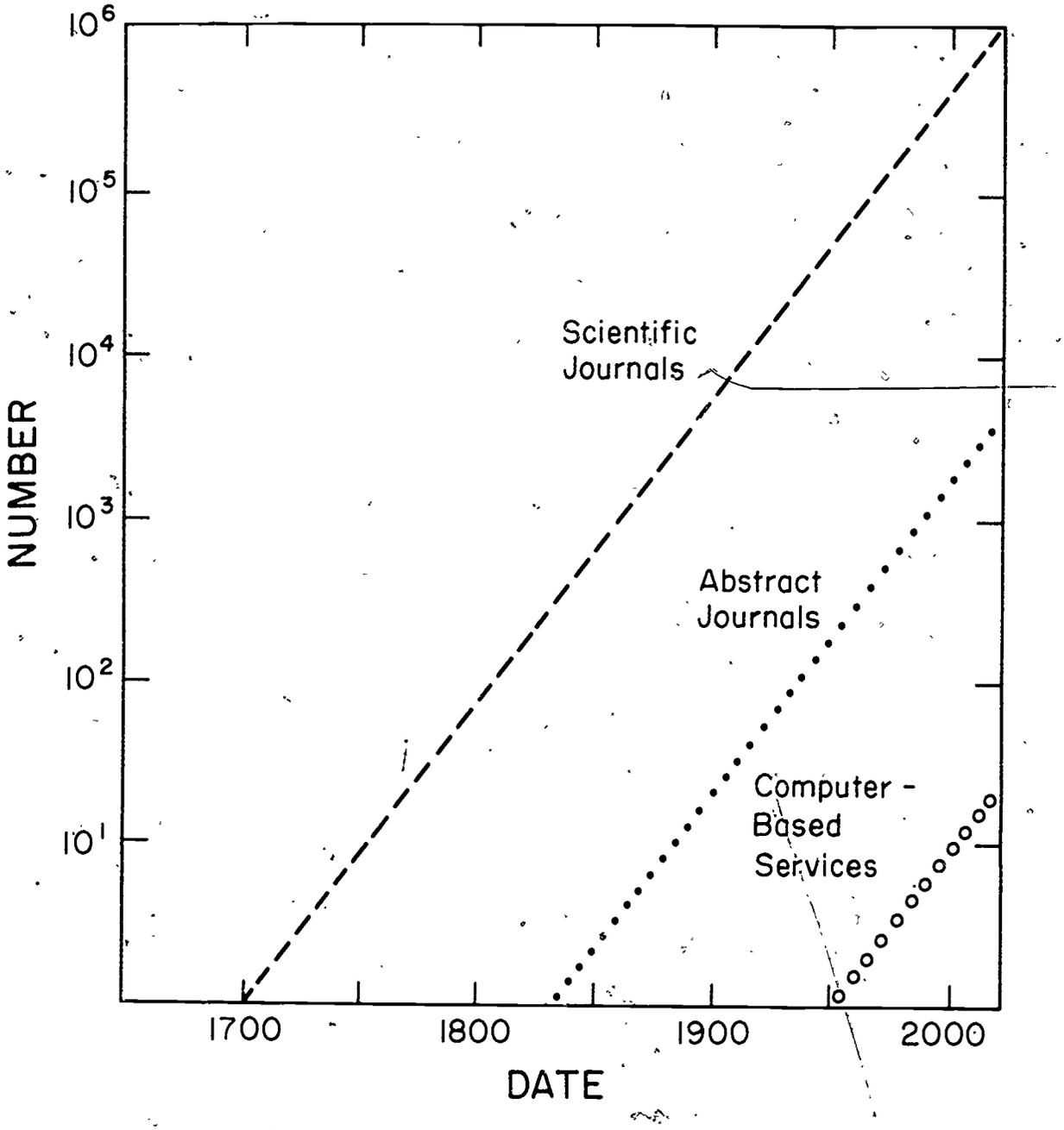


Figure 2.2 Third level of experential representation hypothesized to parallel the growth in primary and secondary literature (after Figure 2.1).

> CA publishes informative abstracts which contain the significant content of published works. A CA informative abstract is a concise rendition of the significant content of a bibliographically cited paper or report which provides enough of the new information contained in the work with sufficient abbreviated details to enable a reader to determine if it is necessary to consult the complete work.

This description of an "informative abstract" leaves a great deal to the discretion of the abstractor, but the publication also gives additional guidance on significant content:

> The following components are considered to be significant in the contents of an article and are included in CA informative abstracts:
> - The purpose and scope of the work, if it is not evident from the title.
> - New reactions, compounds, materials, techniques, procedures, apparatus, data, concepts, and theories.
> - New applications of established knowledge.
> - The results of the investigation.
> - The author's interpretation of the results and his conclusions derived from them.

Wyllys (9) has provided a more succinct, if not more precise, description of the term "abstract."

> An abstract is
> a) a description of, or restatement of, the essential content of a document
> b) which is phrased in complete sentences (except for bibliographic data)
> c) and which usually has a length in the range from 50 to 500 words.

Such descriptions of abstracts as those given above, although wanting in definitional precision, do provide a basis for the formulation of a definition. Thus, the abstract should consist of

complete sentences rather than keywords or phrases, or association maps (10). The abstract should (usually) be short, although the relation between the length of the article and its abstract is prescribed in many different ways (11). The essential content of the article as reflected in the purpose of the work, the results and conclusions, new data, etc., should be included in the abstract. The problem with this last statement is that of determining what is the "essential content"; it is a problem central to all of language processing.

There are two ways of viewing an abstract, as a structural element to be formed and manipulated or as an intensional element. Most, if not all, current attempts at a definition of the term "abstract" mix these two views and consequently fail to provide a definition.

The following definition of an abstract (12), based upon structural considerations, is offered which provides a workable and useful basis for a better understanding of the abstracting process (but, as will be seen, it is incomplete in the form given here).

An abstract, A, is a set of sentences, s, such that

$$A = \{ s \mid s \in D \}$$

where D is the document[6] abstracted, and such that certain transformations on the set, s, are allowed:

concatenation

truncation

phrase deletion

---

[6] Document, as well as other fundamental terms, are defined in (13).

voice transformation

paraphrase

division

word deletion

What does this definition do for us? First, it provides a more precise means of determining whether a document is an abstract, than do earlier descriptions. Thus, one of its most important advantages is that it distinguishes abstracts from critical reviews[7] or other literary inventions of the abstractor. Second, it makes no mention of "content" or "meaning," notions which are difficult to deal with and which may vary with the purpose of an abstract. Thus, the definition distinguishes between what an abstract is and what an abstract does. Third, the definition allows for a certain stylistic freedom, but stylistic freedom does not encompass editorial comment (the latter is precluded by the definition). Finally, the definition is applicable to human, as well as to other, abstractors, thus permitting comparison of the abstracts produced by different abstracting systems.

It should be noted that the definition given above does not prescribe the length of an abstract. Length is a function of the purpose of an abstract and is therefore not of the essence of the concept "abstract." Abstract length is a variable under the control of the abstracting system rather than one whose values are dictated by the

---

[7] An author of a critical review includes added interpretation and criticism of the article that he is reviewing.

definition of the product of the system.

While the definition provided above is useful, it is really incomplete because the intension of an abstract is not provided for. Abstracts are usually placed into one or the other of two intensional classes: informative or indicative. An informative abstract is one characterized as containing some (or all) of the data contained in the original document. An indicative abstract is one which indicates that certain data is contained in the original document, but that data is not contained in the abstract.[8] Examples of informative and indicative abstracts are given in Figure 2.3. Many variations on these two classes of abstract have been described (14). In practice, most abstracts prove to be both informative and indicative, so it is perhaps less important to consider abstracts as belonging to one or the other of these classes than it is to consider the user population they are to serve.

To know the needs and desires of a system's user population is no easy accomplishment. But if I assume such knowledge for existing abstracting services, I can observe that Chemical Abstracts, for example, gives greater emphasis to current work in chemistry than to articles relating historical data (state-of-the-art reviews, books, etc.) because articles of the first type are represented by informative abstracts while the latter type are represented by indicative abstracts. This observation leads to the inference that the users of Chemical

---

[8] Hence an informative abstract is also indicative, but the inverse of this statement is not true.

ABC is judged a more satisfactory process for the water-proofing of synthetic textiles than XYZ. The ABC process yields a product of 20 percent greater durability as judged by standard test #1234. It also yields a better appearing product based on the votes of a panel of 25 textile finishing specialists. The cost of ABC is claimed to be 10 per cent less per square yard although specific cost data are not given. The two processes are about equal in processing speed.

The finishing of textiles by process ABC to achieve water repellency is considered superior to finishing by process XYZ. Factors considered include durability, appearance, cost, and speed.

Figure 2.3 Examples of informative (top) and indicative (bottom) abstracts. (Reproduced from Abstracting Scientific and Technical Literature (10).)

Abstracts are, on the whole, more concerned with current work than with historical data.

These observations lead me to the conclusion that the purpose (which I have inferred from the examination of the products of various abstracting systems) of the abstract dictates the particular set of sentences which constitute the abstract. The purpose of an abstract is met by controlling the intension of the abstract. Thus, a chemical compound appearing in an abstract in Index Chemicus will almost invariably carry with it the intension of "newness," while a compound appearing in Biological Abstracts will most likely bear the intension of application in some biological system, regardless of its "newness." I see, therefore, that the structure of the abstract and the intension of the abstract are not independent, so the definition given earlier must be modified to account for the relationship between structure and intension.

1.4. Formal Definition of "Abstract"

Let me rephrase the definition of abstract given earlier, this time in the language of automata theory. The definition of an abstract can be given in terms of an automaton, M_a, denoted

$$M_a = (K, \Sigma, \Gamma, \delta, q_o, F) \quad (15)$$

where

K is a finite set of states.

Γ is a finite set of allowable input and output symbols: the original document, the abstract, and any additional data.

Σ is the set of input symbols, i.e., the original document.

δ is the next move function which is defined by machine configurations, selection rules, and transformations. This function also specifies the output of the automaton, the elements of the abstract.

q_0 is the start state, which corresponds to the input of the title.

F is the final state, which corresponds to the completion of the abstract.

Although the above definition of "abstract" may seem somewhat abstruse, it is really an operational definition. This definition provides the relationship between the abstract and the original document in terms of an abstracting algorithm. The set of states and associated mappings constitute an algorithm which is a realization of the intention of the machine. The machine can be given explicit definition for a particular abstracting system by specifying all the parameters necessary for operation of the automaton.

Based on the values of certain parameters, various types of abstracts can be defined. When the set of allowable input and output symbols, Γ, contains only the sentences of an original document then the resultant abstract is a selection of sentences from the original, or an extract. When the set Γ contains additional symbols, such as alternative sentence structures and vocabulary items, which allow for the abstract to contain paraphrases of the original sentences, then an informative abstract is produced. When information about the

original document is supplied by the abstracting system, i.e., when such information can serve as input to the automaton, then an indicative abstract is produced. Other types of abstracts that have been identified by authors, for example, alerting abstracts (11) and critical abstracts (14), reflect the orientation of the information that is added to the set of allowable input symbols. By completely specifying all of the parameters in the formulation, an abstract will be completely defined and an algorithm for production of the abstract will also be available.

I have dealt with the concept of an abstract, considering in particular

1. what an abstract is,
2. the purpose of an abstract, and
3. to some extent how an abstract may be produced.

In the following sections, I consider the last point in greater detail. In general, an abstract is produced by an abstractor or an abstracting system. This latter term is more general (the former term seems to imply a human as the abstracting system) and can be applied with equal ease to humans or other kinds of machines which are capable of executing an abstracting algorithm. In the sequel I will briefly consider human abstracting systems and then discuss computer-based systems.

2. Methods of Producing Abstracts

2.1. Human Abstracting Systems

Although automatic abstracting is a popular current topic, all operational abstracting systems are largely human abstracting systems. That is, the reading of the original document, selection of portions of it for the abstract, the writing of the abstract and, in many instances, its formatting and editing, are processes performed by humans. The most important of these processes, from the view-point of the abstract user, is the selection of material from the original document for inclusion in the abstract. But this selection may be influenced considerably by the other processes.[9]

The process of translation from one language to another, as carried out by human abstractors, may result in considerable variation in the size and quality of the abstract. Such variations depend largely upon the abstractor's facility with both the source language and the target language. Although no data are available to support this contention, it seems unlikely that an abstract prepared by an abstractor, whose facility with one of the languages is poor, will be of as good quality (as amplified below) as that prepared by one fluent in that language.

---

[9] A good deal of the discussion to follow is of necessity conjecture, but it should not be particularly difficult (however time consuming) to obtain data with which to affirm or deny the assertions made here. These studies would of course, involve human subjects and unless the experiments were carefully designed and executed, the results would be of doubtful value.

The fault is most serious when the source language is not well understood by the abstractor. Then the abstract may be well written but may not represent the content of the original document. On the other hand, when the target language is the source of difficulty the abstract will show this and the reader will be altered to the possibility for error or misinterpretation on the part of the abstractor.

A second factor which influences the selection of material for the abstract is the abstractor's knowledge of the subject area of the document being abstracted. The abstractor who is expert in the subject area will likely produce an abstract which is shorter, more general and which requires more knowledge on the part of the reader than an abstract produced by an abstractor whose knowledge of the subject area is marginal. On the other hand, the latter type of abstractor is perhaps less likely to get the main thrust of the document into his abstract than the subject expert. What is probably needed is an abstractor whose qualifications lie between the extremes of expertise and passing knowledge in a subject area. Again, there are no data to support these contentions, but the following statement from an editor of Chemical Abstracts (16) lends some credence to them:

> ...the best way to get all of the new and significant information of a paper into an abstract, expressed in proper technical terms, is to obtain the spare-time abstracting service of someone actively interested and working in the specific field of chemistry into which a paper being abstracted fits.

The suggestion is that it is better (and easier) to teach subject specialists how to abstract rather than to attempt to teach abstractors a subject specialty.

Another factor which must influence selection of material for the abstract is the imposition of formatting and editorial responsibilities upon the abstractor. The more mechanical tasks the abstractor must perform, the greater is the likelihood that the quality of the abstract will suffer. Undoubtedly, Zipf's law of least effort (17) comes into play here.

Thus it is suggested that a human abstractor will produce the best abstracts who

1. is expert in the subject area and is taught how to prepare abstracts
2. is fluent in both the source and target languages
3. is subjected to as few purely mechanical tasks as possible.

Although these are not the only factors which influence the quality of human-produced abstracts, they are probably the most influential. These and other factors are considered again later in connection with studies of abstractor consistency.

2.1.1. Operational Systems

Computer systems are presently being utilized in operational abstracting systems to speed up the production of abstract journals. The basic processes in the production of abstract publications can be defined in terms of the following five steps: 1) document selection and acquisition, 2) input processing, 3) abstracting, 4) publication, and 5) announcement and distribution (see Figure 2.4).

The acquisition and selection of documents to be abstracted is essentially a manual operation at present. Most primary sources of

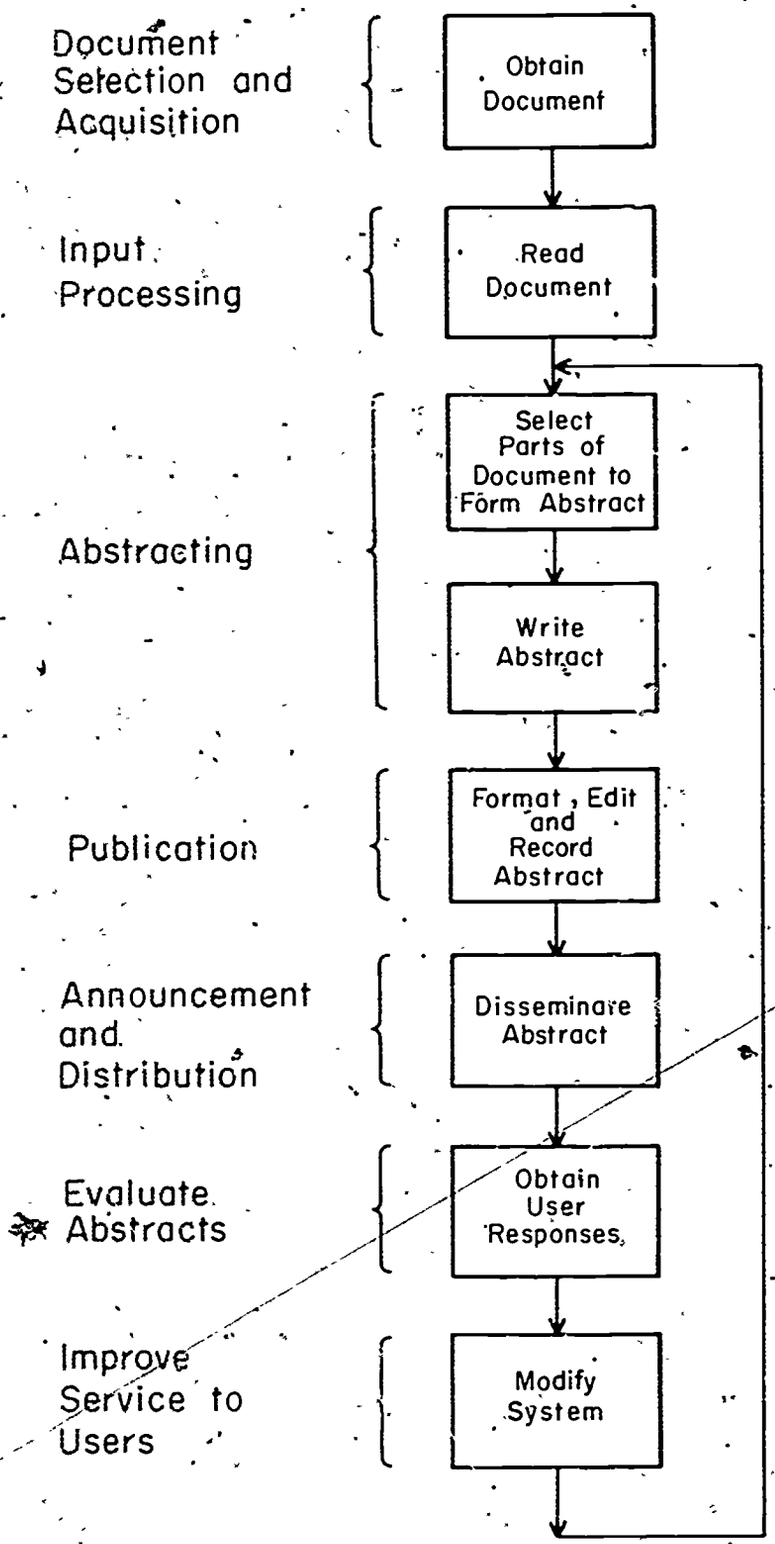


Figure 2.4 Basic processes in the production of abstract publications.

information are received as hard copy and must be converted into machine-readable form for input to the computer system. But current trends in information transfer suggest that increasing reliance will be placed on the exchange of information in machine-readable form. Through use of machine-readable records the process of producing secondary publications can become an integral part of other processes involved in information-transfer technologies. Publishers of primary journals, for instance, may be able to provide abstracting services with the machine-readable data used to typeset the original article.

Source-data preparation, or input processing, today is largely manual and consists basically of keyboarding punched paper tapes, punched cards, magnetic tapes, or disks. This input can then be edited using display-and-edit programs available through cathode-ray-tube terminals. Increased use of optical character readers may also alleviate the need for transcription from hard copy to machine readable form. Furthermore, the availability of primary sources in machine-readable form may eliminate the need for input processing in the foreseeable future.

Abstracting is currently a manual operation in all production environments. Subject experts are employed who have been trained to write abstracts. In order to meet the recognized need for qualified personnel to write abstracts, some journals now require that authors submit abstracts in addition to their articles. This is not a totally adequate solution because the author-generated abstracts are inconsistent with respect to style and coverage and do not, frequently,

reflect the point of view of the abstracting service. Although research efforts appear promising, abstracts which are produced by computer can not yet meet the demands of abstract-journal publications. The abstracting operation itself will probably be the last to be fully automated in the production of abstracts.

Many abstract publications rely heavily on automatic data processing equipment for production of their journals. Secondary publications, such as Index Chemicus, Biological Abstracts, Psychological Abstracts, Index Medicus, and others like them, have pioneered in the use of computer-aided composition and typesetting. In most applications, the information handling capabilities of computers in printing have been used primarily for right-margin justification, syllabification, and page composition or for instructing hot metal or photocomposition machines and teletransmission devices. Such modern typographic and printing tools as photo-composition, xerographic processes, and page composition, will certainly continue to make printing faster, if not more economical (18).

Announcement and distribution of secondary services is being enchanced by the development and use of computer-based selective-dissemination-of-information systems. The selective dissemination of information to users is made on the basis of user profiles, each of which is a compilation of keywords reflecting a user's interests. The user is then presented with only those abstracts which have matched his profile. The goal of such systems is to present to the user only the abstracts which are of the greatest potential value to him.

Chemical Abstracts Service[10], long a leader in improvements in abstract journal publication, recognized in the early 1960's that the traditional manual system for processing and publishing secondary chemical information was too slow, too expensive, too rigid, and too wasteful in its use of highly capable manpower to be effective in the face of the ever growing volume of published information. A streamlined system was designed in which each primary document would be analyzed only once and the selected data recorded only once in a form that could be used to produce a variety of information packages with overlapping content. The target system designed by the Chemical Abstracts Service for their operations combines human intellectual analysis and computer-based processing. The roles of the computer in the system are:

1. To receive material derived by human intellectual analysis;
2. To support that analysis with machine aids and augment the information flow by retrieving related previous work;
3. To apply automated validation checks and trigger exception reviews by editorial staff;
4. To eliminate the necessity for manual bridging between processing steps;
5. To automate the ordering (sorting) and formatting of the information, both on a data-directed basis;
6. To control composition machinery; and

---

10 Other publications, such as Physics Abstracts, Biological Abstracts, and Psychological Abstracts are also moving toward the incorporation of sophisticated man/machine systems in their production.

7. To provide computer-readable files.

This target system, which was initiated during the 1967-68 period and will probably not be completed until the 1975-76 period, represents a system which uses computer technology as an essential and integral component (19).

2.1.2. Selection and Consistency

Before any attempt is made to automate the process of abstracting, it is important to understand how humans produce abstracts. A number of studies have been reported which deal with the question: How does a human abstractor decide what material in the original document should be included in the abstract? A corollary question is: Does the abstractor display any consistency in the selection process

a. Within a document;

b. Between documents;

c. Through time for either a. or b.?

In order to answer these questions, one needs to know

1. What constitutes a good abstract of a document (or what constitutes a good set of abstracts, if one allows for the possibility of several equally good, goal-directed abstracts);

2. What is the operational definition of an abstract used by a particular abstracting system, as embodied in the rules for abstracting prescribed (or described) by the system[11] and

---

11 The directions for preparing author abstracts, issued by a particular journal constitute an operational definition (no matter how vague) of an abstract.

how does it contrast with the definition of a "good" abstract;

3. Given answers to 1 and/or 2, what consistency does an abstractor display in following the rules of the abstracting system or in producing a good abstract (if the rules of the abstracting system seem, or are believed to be, at odds with the abstractors' understanding of what constitutes a good abstract).

These three questions can be studied alone or in pairs, as illustrated in Figure 2.5. Only the last of these questions has been studied systematically, yet there is no clear answer to any of these questions, although some of the research which has been done at least suggests the direction further efforts should take. Such studies have a special significance for the development of automatic abstracting systems, as will be made clear subsequently.

The most significant studies on selection and consistency are those of Rath, Resnick and Savage (20, 21, 22, 23) and of the Thompson Ramo Wooldridge, Inc. (TRW) group headed by Edmundson and Wyllys (24, 25).

The main conclusion to be drawn from the Rath-Resnick-Savage studies is that human abstractors show poor consistency in selecting sentences for the abstract both between abstractors and with respect to time. In an initial study (22), these workers assigned 6 human abstractors the task of reading a set of 10 articles from Scientific American, choosing the 20 most representative sentences from each article, and ranking these sentences in decreasing order of representativeness (measured against a background of those sentences already

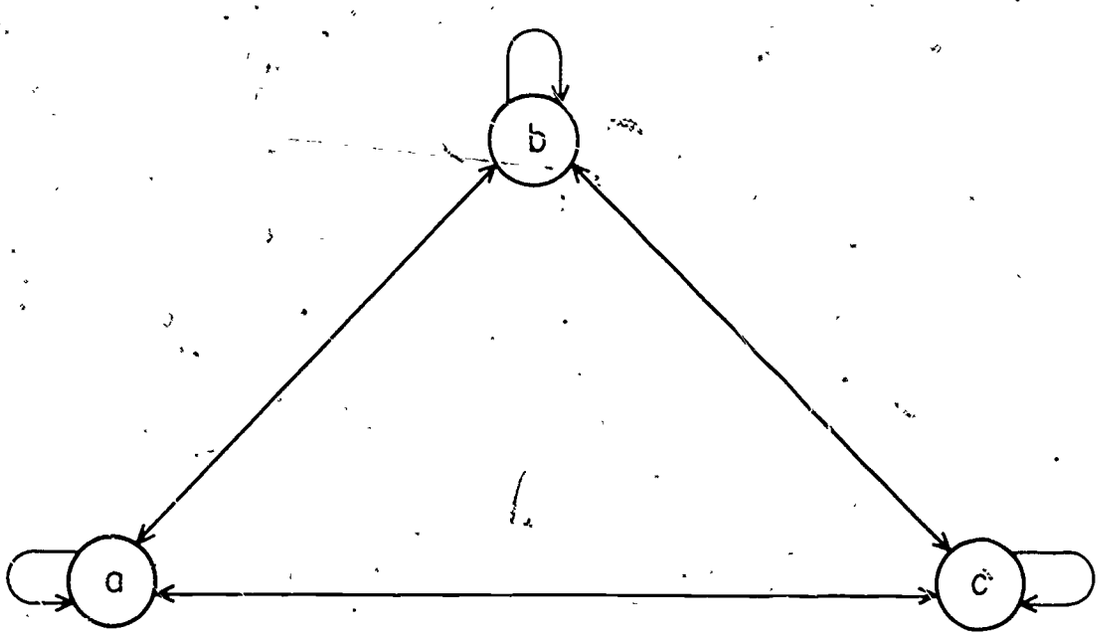


Figure 2.5 Pairwise combinations of questions in the study of sentence selection and consistency of abstractors. The questions to be studied are: Does the abstractor display any consistency in the selection process

a. Within a document,
b. Between documents, and
c. Through time for either a or b?

ranked). At the same time a computer program, similar to one developed by Luhn (26) (see Section 2.2.2.1) was used to select and rank 20 sentences from each of the test articles. Five different methods of ranking sentences were used in the computer-based abstracting procedure.[12] One interesting point concerning what sentences were selected by human and computer was reported. It was found that the 10 articles taken as a whole contained 37% "topic sentences" (after Baxendale (27)). Of these topic sentences, humans selected 47% and the computer selected 33%. It was also found that human-selected sentences came more often from the first half of the article, while the computer made more sentence selections from the latter half of the article.[13]

In a follow-up study, Resnick (38) showed that abstractors differed in their consistency in abstracting the same article at fairly widely separated points in time. Five subjects were asked to abstract 6 articles from Scientific American, as in the previous study (23). Eight weeks later, they were asked to abstract the same six articles. Instructions included the admonition not to select sentences which had previously been selected unless the subjects felt the sentences were still representative, and to mark any currently-selected sentence they

---

[12] It is important to note that the sample sizes used in this experiment are much too small to yield statistically valid results, although the data do support what one would feel, intuitively, was true. Nevertheless, the questions which this study attempted to answer remain open.

[13] This result may suggest that human abstractors get tired more readily than computers. Furthermore, the conditions under which the subject-set of human abstractors worked was probably a poor approximation of the (often harried) conditions under which the professional abstractor works.

believed they had previously selected. The accuracy of these identifications is indicated in Table 2.1. On average, these abstractors were able to correctly identify a currently-selected sentence as one which had been selected previously 42.5% of the time.[14]

Table 2.2 gives data on the consistency in sentence selection over time. It can be seen that there is greater variation among abstractors than between articles for a given abstractor.[15] One may conclude, from these studies (keeping in mind the smallness of the samples), that human abstractors are modestly consistent in producing abstracts of a given document.

In another study of abstractor consistency, carried out by the TRW group (25), it was concluded that although human abstractors were not very consistent among themselves, the abstracts they produced were adequate (in terms of inter-abstractor consistency) to justify a study of the attributes of the sentences they extracted from the document.

The way in which the TRW group measured the consistency of human abstractors is of some interest. A correlation coefficient was devised which attempted to measure the similarity between two sets of sentences

---

[14] This result suggests that the subjects' memories functioned less well than their abstracting algorithms.

[15] It would be interesting to know the reason for the individual variations, which deviated widely from the average for a given abstractor. In Table 2.2 abstractors 2 and 4 (particularly 4) show between-article variations which are inconsistent with these variations for the other subjects.

Table 2.1 Mean percentage of responses of five subjects making selections of 20 "representative" sentences two months apart from each of six articles. During the second selection the subjects were asked to indicate whether or not they had chosen each sentence two months earlier (from (23)).

	Sentence correctly identified	Sentence incorrectly identified
Sentence previously selected	42.5%	13.1%
Sentence not previously selected	21.7%	22.7%

Table 2.2 Percentage of sentences selected on second trial which were the same as those selected on the first trial (From (23).)

	Articles					
Abstractor	A	B	C	D	E	F
1	60%	55%	45%	45%	40%	40%
2	45	50	75	80	45	60
3	60	65	55	70	55	50
4	45	55	55	55	40	15
5	80	70	70	60	55	80

extracted from a given document, taking into account the sizes of the sentence sets relative to the original document and relative to each other. This measure can be expressed as

$$q_j = \frac{S_j}{L_j}$$

where L_j is the length of the sentence T_j (number of text tokens) and

$$s_j = \sum_{\forall w_i \in T_j} P_i$$

where

$$P_i = \frac{N_{w_i}}{M_{w_i}}$$

where N is the number of text tokens of type w_i in the sentences extracted from a set of documents and M is the number of text tokens of type w_i in the document set ($M > N$). The probability that a sentence will be extracted given N, M, L and w_i is q_j. This probability can be derived from a consideration of the concept of inductive probability (developed by Carnap (29)) as follows (25).

Suppose two gamblers X and Y are told that a certain document has been extracted and that a particular sentence of that document has certain specified characteristics. These characteristics, together with all data available on past sentence selection by abstractors, are summarized in a proposition called the evidence, E. Let S be the

hypothesis that the particular sentence has been extracted. Let the betting quotient, q, for a bet on S be the ratio

$$q = \frac{\text{stake offered for a bet on S}}{\text{total stake}}$$

Suppose the amount bet on S is q and the amount bet against S is 1-q. Then one can assume the existence of some q such that X would be equally willing to bet for or against S. Such a value of q is, as far as X is concerned, the psychologically fair betting quotient for S relative to the evidence, E, available. Another way of expressing the meaning of q is to say that q indicates the preference that X believes E confers on S. Y may be assumed also to accept either betting role for a q having the above value.

To determine whether q is indeed fair, a given document should be abstracted by a large number of abstractors.[16] If X and Y made a series of bets, X for S with quotient q and Y against S with quotient 1-q, then the total balance of wins and losses would be zero if the ratio of S_{true} to S_{false} is exactly q.

The TRW group (24) concluded that a representativeness score for a sentence under hypothesis S should incorporate:

1) the degree of evidential support that E confers on S;

2) the fair betting quotient for S relative to E;

---

16 Such a determination is always hindered, in practice, by the practical difficulties of carrying out a study involving large numbers of abstractors and/or large numbers of documents.

3) an E-based estimate of the ratio

$$\frac{S_{true}}{S_{false}}$$

This measure is a general one because it permits the evidence, E, to be specified in any desired manner. Thus, although the TRW study used the measure as support for sentence selection based essentially on frequency criteria, any well-defined criteria may be used. The important point is that if the fair betting quotient q can be determined, then the selection criteria may be incorporated in a program for automatic abstracting.

In concluding this section, it is fair to say that little is known about how or why human abstractors choose from the original article what they include in the abstracts which they produce. Neither is it clear to what extent human abstractors are consistent in abstract production. Perhaps more importantly, especially since there is no concrete answer to the question of what constitutes a good abstract, questions relating to human selection and consistency may be irrelevant. It is just possible that the abstracts produced by humans are not good. If so, then it would be undesirable to try to emulate by computer the processes which humans use in abstracting, since such emulation would lead simply to a faster rate of production of consistently poor abstracts.[17] The major, unanswered question is:

---

[17] A product easy to produce and difficult to sell.

"What makes an abstract a good one?"

2.2. Computer-Based Abstracting Systems

I turn now to a discussion of abstracting systems in which the computer plays a central role. This discussion must, like that for human abstracting systems, be tempered by a lack of knowledge of whether the abstracts produced by such systems are "good." Happily, however, the questions of selection and of consistency of selection are easily dealt with. It is these two aspects of computer-based abstracting systems[18] that will be considered in greatest detail in this section.

There are eight[19] significant studies related to computer-based abstracting which have been described in the literature. These are

1. The Luhn study;
2. The ACSI-matic study conducted by IBM;
3. The Oswald study;
4. Word-association research;
5. The TRW studies conducted by Edmundson and Wyllys;
6. The Earl study;

---

[18] I prefer to call the process of abstracting by computer "computer-based abstracting"; the abstracting system a "computer-based abstracting system"; and the abstract produced by such a system a "computer-produced abstract." rather than to use the term automatic in place of computer-based as has commonly been done in the literature.

[19] To be sure, one could include other studies in this list, notably those of Baxendale (2) and of Climenson, Hardwick and Jacobson (30), but I believe those listed to be representative of the various approaches taken to the production of abstracts by computer program, and that they are also of some historical importance.

7. The Soviet study;

8. The Rush study.

I will discuss the first seven of these studies in the remaining sections of Chapter II and I will discuss the Rush study in Chapter III. Before I discuss these studies, it will be useful to describe the basic components of a computer-based abstracting system.

2.2.1. Basic System Components

A computer-based abstracting system (see Figure 2.4) must

a. read the document to be abstracted,

b. analyze the document,

c. apply a set of selection and/or transformation rules to produce the abstract,

d. format the resulting abstract, and

e. print the abstract.

Reading[20] the original document is perhaps as difficult a task as any subsequent processing steps the abstracting system must perform. A technical paper will commonly use many different characters in several different sizes and in a half-dozen different fonts. The ordinary computer-system input devices are not capable of handling such a wide range of characters, nor are there optical readers (optical scanning devices) available which can do so economically.[21] One or

---

[20] Reading obviously involves seeing, and there are a variety of methods for recording data so that a computer can "see" (sense) it. These methods are beyond the scope of this work, but the reader is referred to (31, 32) for details and further references.

[21] Except, perhaps on a very large scale.

both of two simple solutions to this problem are usually effected: 1) pre-edit the text, eliminating or altering those portions of the document which the input devices cannot handle; 2) use a scheme of flagging (coding) of characters which cannot be read directly. Figures, tables and other graphic materials will in all probability not find their way into the computer system at all. Pre-editing is most likely to be used, since the special text features which might otherwise be preserved are of questionable value in the actual abstracting process and since, until the use of optical printers becomes wide-spread, computer printers cannot economically (if at all) print the necessary range of characters.

Once the text has been read into computer memory, analysis of the input text is performed. The actual analytical methods used depend upon the particular abstracting system, so a discussion of this component will be deferred until section 2.2.2.1.

Similarly, rules for selecting parts of the original document for inclusion in the abstract are considered under each of the specific studies which are discussed, beginning in section 2.2.2.1.

Formatting the output of the abstracting process has not been particularly inspired. The usual methods include listing individual sentences or, occasionally, printing the abstract paragraphed as was the original document. Abstracts have never been noteworthy examples of literary work nor of the typographic art, so one should not criticize the output of computer-based abstracting systems too severely (although it does not seem unreasonable to hope that a little imagination might be applied to the formatting of computer-produced

abstracts). No existing computer-based abstracting system has utilized any output device other than the line-printer.[22]

2.2.2. Major Research Efforts in the Production of Abstracts

With this general view of the basic processes involved in computer-based abstracting, I now turn to a discussion of specific studies which have lead to the implementation, at least experimentally, of computer-based abstracting systems.

2.2.2.1. The Luhn Study

Luhn is credited with having first suggested and demonstrated that abstracts could be produced via computers (26). The procedures employed by Luhn for generating abstracts by computer were as follows.

a. The document to be abstracted was first punched into cards (texts which were used required no pre-editing) and then transfered to magnetic tape.

b. The text was then read, word by word. Common words[23] were deleted through table look-up. The remaining words, called content words, were associated with any punctuation that

---

22 Certain abstract journals are printed in a two-step process in the first of which an optical printer (COM, computer-onto-microfilm processor) is used. These publications do not, however, result from the compilation of computer-produced abstracts.

23 Common words might well be called non-substantive words, since these are words that are considered to have no value in determining the significance of a portion (sentence, paragraph, etc.) of text. Common should not be confused with function in reference to word classification.

preceded or followed them, and their exact location in the original document was noted.

c. The content words were then sorted into alphabetical order.

d. Words of similar spelling were "consolidated" as follows. Successive pairs of word-tokens[24] were compared letter-by-letter and at the first point of difference a count was initiated of the number of non-matches observed from that point to the end of the longer word-token. If this count was less than seven (<7), the word-tokens were taken to be of the same word-type (i.e., to represent the same notion), otherwise the word-tokens were taken to be distinct word-types. The frequency of occurrence of each word-type was then determined and word-types occurring less frequently than some prescribed value were deleted. The remaining word-types were considered to be "significant".

e. The significant word-types were then sorted into location order.

f. Sentence representativeness was next determined. Sentences were divided into substrings each of which was bounded by significant words separated by no more than four non-significant words. (Significant words separated from other significant words by more than four words were called "isolated" words

---

[24] A word-token (sometimes text-token) is a place-holder in a text. Each different word-token is called a word-type (or text-type).

and were not given further consideration.) For each substring, a representativeness value r_i, was calculated according to the equation

$$r_i = \frac{p_i}{q_i}$$

where p_i is the number of representative tokens in the cluster and q_i is the total number of tokens in the cluster. The highest r_i for a sentence is taken as its representativeness of document content. Sentences having a value of r_i above a prescribed value (or else a predetermined number of sentences of highest r_i) were selected for inclusion in the abstract.

g. The abstract was then printed (as a set of sentences), formatted in paragraph style.

While the methods employed by Luhn for determining word and sentence significance have fallen somewhat into disrepute, the technique is clearly of historical importance. An example of an abstract produced by this method is given in Figure 2.6. Additional examples of abstracts produced by this method, as well as data on word representativeness and word-token counts may be found in Luhn's papers (26, 33, 34).

2.2.2.2. The ACSI-Matic Study

The ACSI-Matic study, conducted by IBM (35, 36) for the Army Department's Assistant Chief of Staff for Intelligence, was (and is) the only study which lead to a computer-based abstracting system as part of an operational information system.

Exhibit 1

Source: The Scientific American, *Vol: 196, No. 2, 86-94, February, 1957*

Title: *Messengers of the Nervous System*

Author: *Amodeo S. Marrazzi*

Editor's Sub-heading: The internal communication of the body is mediated by chemicals as well as by nerve impulses. Study of their interaction has developed important leads to the understanding and therapy of mental illness.

Auto-Abstract*—Exhibit 1

*It seems reasonable to credit the single-celled organisms also with a system of chemical communication by diffusion of stimulating substances through the cell, and these correspond to the chemical messengers (e.g., hormones) that carry stimuli from cell to cell in the more complex organisms. (7.0)***

Finally, in the vertebrate animals there are special glands (e.g., the adrenals) for producing chemical messengers, and the nervous and chemical communication systems are intertwined: for instance, release of adrenalin by the adrenal gland is subject to control both by nerve impulses and by chemicals brought to the gland by the blood. (6.4)

The experiments clearly demonstrated that acetylcholine (and related substances) and adrenalin (and its relatives) exert opposing actions which maintain a balanced regulation of the transmission of nerve impulses. (6.3)

It is reasonable to suppose that the tranquilizing drugs counteract the inhibitory effect of excessive adrenalin or serotonin or some related inhibitor in the human nervous system. (7.3)

* Sentences selected by means of statistical analysis as having a degree of significance of 6 and over.
** Significance factor is given at the end of each sentence.

Figure 2.6 Example of an abstract produced by Luhn's system. (Reproduced from The IBM Journal of Research and Development).

ACSI-Matic employed selection procedures analogous to those suggested by Luhn; the importance of this study lies in the rather novel variations imposed on Luhn's basic techniques. These modifications included:

a. elaboration of Luhn's sentence scoring technique

b. special treatment of documents with an unusually large fraction of low-frequency words

c. special treatment of documents with extraordinarily long sentences

d. choice of sentences, once scored, to form a tentative abstract

e. procedures to reduce redundancy among the sentences selected.

The ACSI-Matic scoring technique improved upon Luhn's treatment of the density of representative words in a sentence. To illustrate the technique, consider the sentence

N R N N R N N N R N N

where N = a nonrepresentative word (N-words) and R = a representative word (R-words). R-words were given a value of 1 and non-terminal sequences of N-words were given a value of $1/2^n$, where n is the number of N-words between successive R-words. Thus, the sentence above would be scored

$$1 + 1/4 + 1 + 1/8 + 1 = 3+3/8.$$

This procedure was applied to sentences of documents whose average sentence length was in the range 18 to 26 words.

When a document had an average sentence length greater than 26 words, each sentence score, computed as above, was divided by the square root of the number of words in the sentence to give a "corrected" score. The procedure favored slightly selection of longer sentences. On the other hand, when more than 10% of the sentences of a document exceeded 40 words in length, the unmodified scoring procedure was employed once again. Thus, the overall effect of these procedures was to give a slight preference for selection of sentences whose lengths were in the range 26 - 40 words (other things being equal).

The above scoring techniques were based on the assumption that words whose frequency exceeded the average word frequency within the document were "representative". This assumption was made when 48% to 56% of a document consisted of function words.[25] When the percent of function words in a document fell outside this range, special treatment of the document was effected. When there were more than 56% function words in a document, the list of potentially representative words was reduced by deleting all words whose frequency of occurrence was greater than 1% of the word-token count for the document.

If there were less than 48% of function words in a document and the document contained more than 35% of unit-frequency words, those words were chosen as representative whose frequency of occurrence was

---

25 Function words are those words not included in one of the classes: noun, verb, adjective, adverb. It is emphasized that there must be maintained a clear distinction between function word and non-substantive words.

less than or equal to the average word frequency for the document.

Once all sentences in a document had been scored, a set of sentences which were potentially members of the abstract were selected as follows.

> The number of sentences in the document is divided by ten. If the quotient is more than 20, 20 is subtracted from the result and the remainder is divided by 32. The number of sentences in the abstract is this quotient plus 20. If the document has fewer than 200 sentences, the abstract has 10 percent of the total number of sentences. (36)

The n sentences with highest scores were designated as "abstract sentences" and the n/4 sentences with next highest scores were called "reserve sentences". Word-tokens were consolidated by a process analogous to Luhn's method following which the "abstract sentences" were examined for possible redundancy. Two sentences were considered to be redundant if they contained a number of matching words which was greater than 1/4 the total number of words compared in the two sentences. Highly redundant sentences were deleted and were replaced by sentences from the "reserve sentence" set. This process was continued until no more redundant "abstract sentences" were found or until the set of "reserve sentences" was exhausted. The sentences that were in the set of "abstract sentences" at the end of this process constituted the computer-produced abstract.

The AGSI-Matic study thus employed several interesting criteria for selecting sentences from a document to form an abstract. The computational complexity of the algorithms necessary to perform the tasks outlined above is clearly considerable, and the selection criteria have not been evaluated.

2.2.2.3. The Oswald Study

The essential distinction of the Oswald study (24) is that he employed an indexing criterion in the selection of sentences to form an abstract. This indexing criterion was that groups of words, as well as single words, should be index entries. Consequently, Oswald chose sentences for an abstract that scored high in the number of "representative" word groupings they contained. Representativeness was determined in a manner similar to that employed by Luhn (26).

Oswald's procedure involved the following steps (24).

1. Determine the count of the tokens of only those words which are significant in the content of the document.
2. Next, identify the highest-frequency words and note words adjacent to them which had a frequency of occurrence greater than one. Such juxtaposed words formed "multiterms".
3. Identify those sentences which contain two or more multiterms, rank them in decending order of multiterms and select some number of the highest ranked sentences according to some prescribed criterion for the length of the abstract.

Since Oswald did not have a computer at his disposal, he was obliged to use human simulation of these procedures for his study. And when one considers that the identification of words "significant in the context of the document" (24) entailed a subjective judgement, these procedures could not readily be implemented directly on a computer. Nevertheless, the study is significant for three reasons:

1. for its recognition of the relationship between an index and an abstract;
2. for its realization that an abstract represents, at least in part, an attempt to concentrate the essence of a document as much as possible;
3. for its recognition of the fact that word-strings of length greater than 1 (multiterms) are important in determining sentence significance.

2.2.2.4. Word Association Research

The significance of word-association studies is that they further emphasize the importance of word clusters (a fact already emphasized above) in conveying an author's intent. Doyle (37, 38, 39), Bernier (40), Quillian (41) and others have attempted to represent a measure of semantic information content (i.e., the measure of some textual structure's value in conveying an author's intent) through the use of special representations of language.

Doyle's association map concept is based upon statistical association criteria and a map could, therefore, be produced by computer program. The method for generating an association map involves the creation of an n x n correlation matrix (n = number of key-words to be correlated), and the correlation of the matrix elements, by means of the Pearson correlation coefficient (38), for co-occurrence of key-

words in individual documents.[26] A portion of a typical association map is shown in Figure 2.7. Links between nodes (words) show the word associations. Arrows indicate that word co-occurrence is due mainly to a two-word term; the arrows point to the second word of the term.

Doyle envisioned the use of an association map as an associatively organized index which would facilitate the operation of a combined man-machine searching system. But the association map derived from a single document represents a telegraphic abstract in which an attempt has been made to indicate term relations through statistical association.

Quillian's work is superficially similar to that of Doyle, but in the generation of what one might call "semantic maps" Quillian makes the associations intellectually rather than statistically. These semantic maps constitute a memory model which Quillian later employed in studies of human-like language behavior (41). The significance of this work for abstracting lies in the strong emphasis on relations between words (concepts) rather than on the words alone. Although the details of this work cannot be treated here, the reader is referred to any of several good papers for this purpose (43, 44, 45).

In addition to the works cited, Bernier (40) and more recently Avramescu (46) and Fugmann (47) have emphasized the importance of specifying both words and relations between words in indexes and

---

[26] In one experiment, Doyle (37) employed the document collection and corresponding index terms (keywords) already employed in another connection by Borko (42).

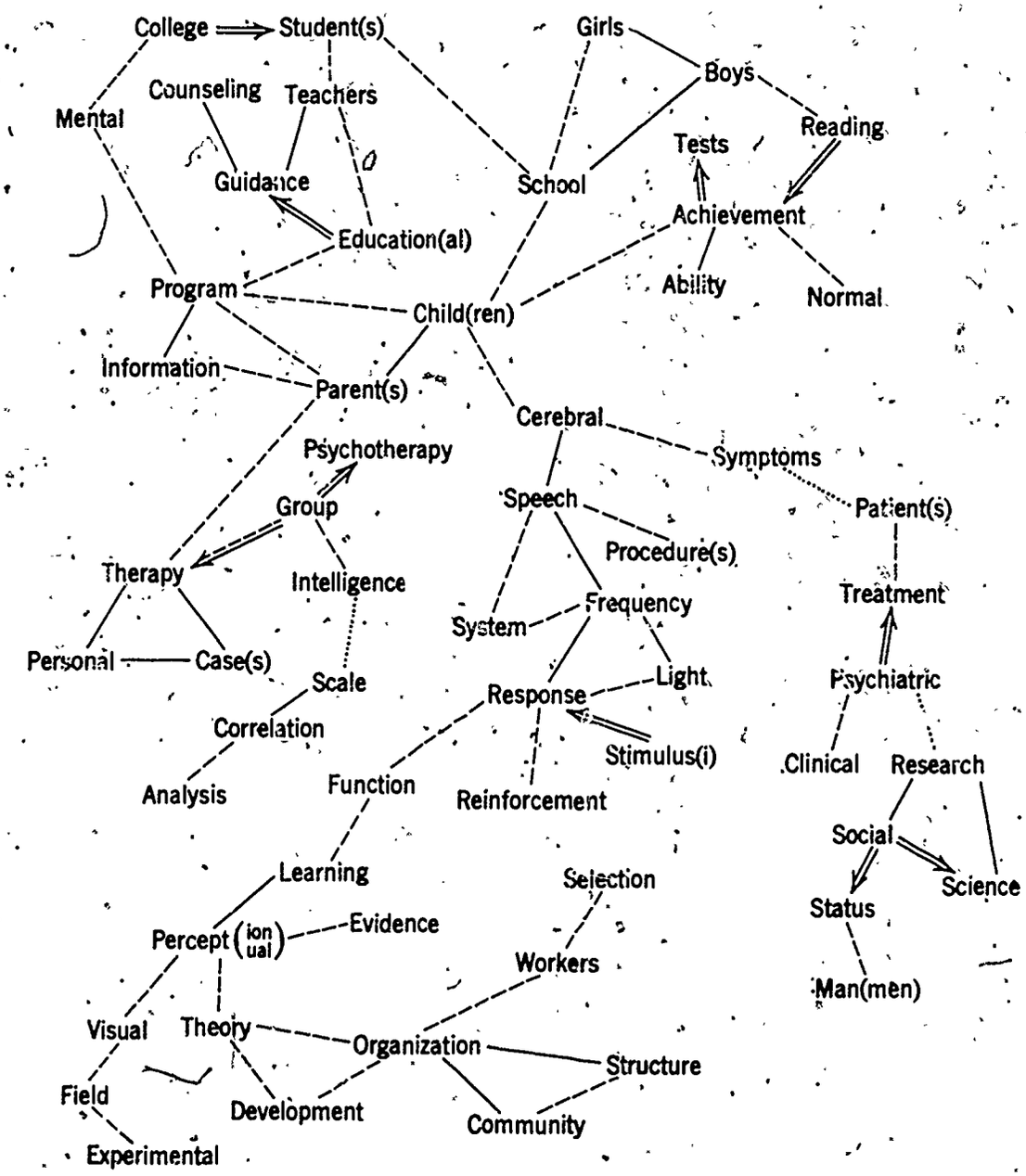


Figure 2.7 Typical association map of Doyle, based on Pearson correlation coefficients. (Reproduced from Automated Language Processing (39).)

abstracts. Nevertheless, there has been no concerted effort to use semantic structure as a basis for creating abstracts of original documents. But since abstracting is fundamentally a matter of semantic content abstraction, it is reasonable to expect that studies along the lines of the investigations mentioned here would be of value in the computer-based production of abstracts.

2.2.2.5. The TRW Study (Edmundson and Wyllys)

The objectives of the Thompson Ramo Wooldridge, Inc. (TRW) study in computer-based abstracting were twofold: the development, first, of an abstracting system to produce indicative abstracts, and second, of a research methodology which would permit new text and new abstracting criteria to be handled efficiently (25, 48, 49, 50, 51, 52, 53, 54). The research methodology comprised a study of the abstracting behavior of humans, a general formulation of the abstracting problem and its elation to the problem of evaluation, a mathematical and logical study of the problem of evaluation, a mathematical and logical study of the problem of assigning numerical weights to sentences, and a set of abstracting experiments employing a cycle of implementation, testing and improvement. The research concerned with the abstracting behavior of humans has already been discussed in section 2.1.2.

The evaluation of the quality of abstracts produced by any system is necessary in order to make improvements in that system. The TRW group considered five methods of evaluation (25).

1. Intuitive value judgement

2. Creation of "ideal" abstracts to serve as standards of comparison
3. Construction of college-type test questions on the document to be answered from abstracts by a sample population (evaluation of the summary function of an abstract)
4. Retrievability of the document via the abstract (evaluation of the retrieval function of an abstract)
5. Statistical correlations (applicable only to extracts)

Method 2 was implemented early in the research with the results giving an indication of how "human-like" the selection of sentences is. The method was not considered a final evaluation of abstracts, but only a rough indication of agreement with human selection of sentences. This method served primarily as a rejection test for abstracts which were little better than random selection of sentences. Evaluation methods 4 and 5 were implemented later in the research to indicate areas of content improvement.

The TRW study developed a logical, mathematical method for the assignment of numerical weights to sentences. This study showed that there is considerable potential in a set of four methods of sentence selection which they called the Cue, Key, Title, and Location methods. These methods can be described, briefly, as follows (54).

The Cue method makes use of a list of words which are classified as bonus words (those that have a positive value, or weight, in sentence selection), stigma words (words that have a negative weight) and null words (those words which are irrelevant to sentence selection).

The Key method is based on the frequency of occurrence of words, and

is similar in approach to the method used by Luhn in his pioneering study.

The Title method is based on a glossary of the words of the title and subtitles (excluding null words). Sentences containing words that also occur in the title are assigned a higher weight either than sentences which contain words that occur also in a subtitle, or sentences which have no such words (other things being equal).

The Location method is based on the hypothesis that certain headings precede important passages and that topic sentences occur early or late in a document or paragraph. This latter was also Baxendale's hypothesis, resulting from observations made in her studies of automatic indexing (27). These four methods are summarized in Figure 2.8.

In the final system the relative weights among the four basic methods were parameterized in terms of the linear function

$$a_1C + a_2K + a_3T + a_4L$$

where a_1, a_2, a_3, and a_4 are the parameters (positive integers) for the Cue, Key, Title, and Location weights, respectively. The mean percentages of coselection of sentences from the ideal abstract and the test document for the most interesting methods are shown in Figure 2.9, with the intervals encompassing the sample mean plus and minus one sample standard deviation. The Cue-Title-Location method is seen to have the highest mean coselection score, while the Key method in isolation is the poorer of the methods. On the basis of these data it was decided to omit the Key method as a component in the preferred abstracting process. An example of an abstract produced by use of the combination of the Cue, Title, and Location methods is shown in

		Structural Sources of Clues:	
		Body of Document (Text)	Skeleton of Document (Title, Headings, Format)
Linguistic Sources of Clues:	General Characteristics of Corpus	CUE METHOD: Cue Dictionary (995 words) (Includes Bonus, Stigma, and Null subdictionaries)	LOCATION METHOD: Heading Dictionary (90 words) (Location method also uses ordinal weights)
	Specific Characteristics of Document	KEY METHOD: Key Glossary	TITLE METHOD: Title Glossary

Figure 2.8 Rationale of the four basic sentence selection methods employed by Edmundson. (Reproduced from The Journal of the Association for Computing Machinery (54).)

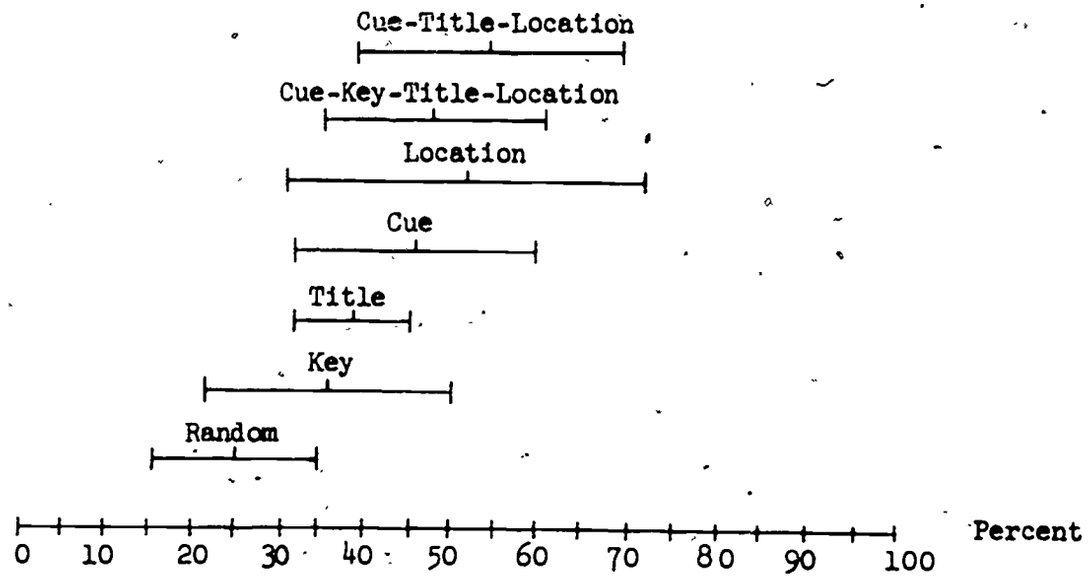


Figure 2.9 Mean coselection scores obtained from the sentence selection methods employed by Edmundson. (Reproduced from The Journal of the Association for Computing Machinery (54).)

Figure 2.10.

The results of the research at TRW indicated that abstracts can be defined, identified and produced in a computer-based system. They concluded that future abstracting methods must take into account syntactic and semantic characteristics of the language and text; they could not have relied simply upon gross statistical evidence. Edmundson concluded that the major task of any further research would be to identify and define the differences between manual and computer-produced abstracts and to minimize these differences so that computer-produced abstracts can supplement, and perhaps compete with, traditional ones (54).

2.2.2.6. The Earl Study

The investigation of computer-based informative abstracting and extracting performed by Earl and her associates at Lockheed Missiles and Space Company has been aimed at basic research in English morphology, phonetics, and syntax (55, 56, 57, 58, 59, 60, 61, 62, 63). This study, which has been supported by the Office of Naval Research since 1964, has dealt with basic linguistic research as a necessary prerequisite to the production of abstracts by computer.

During the first three years of the program, a word-data base was established. This data base along with a part-of-speech algorithm was used to provide an algorithmic determination of the parts of speech of written English words. The parts of speech were later used to determine if there existed any linguistic similarity between sentences that were used for abstracts. Also, during these years, it was

PAR SENT DOCUMENT NUMBER 6 PAGE 1

ABSTRACT BASED ON CUE TITLE LOC. WTS.

EVALUATION OF THE EFFECT OF DIMETHYLAMINE BORINE AND SEVERAL OTHER ADDITIVES ON COMBUSTION STABILITY CHARACTERISTICS OF VARIOUS HYDROCARBON TYPE FUELS IN PHILLIPS MICROBURNER (AOBTT30)

R. L. BRACE

1 0 SUMMARY

2 1 AT THE REQUEST OF THE NAVY BUREAU OF AERONAUTICS, PHILLIPS PETROLEUM COMPANY UNDERTOOK THE EVALUATION OF DIMETHYLAMINE BORINE AS AN ADDITIVE FOR IMPROVING THE COMBUSTION CHARACTERISTICS OF AVIATION GAS TURBINE TYPE FUELS.

2 2 BECAUSE OF THE SMALL AMOUNT (100 GRAMS) OF DIMETHYLAMINE BORINE RECEIVED FROM CALLERY CHEMICAL COMPANY, THIS EVALUATION HAS BEEN LIMITED TO THE MEASUREMENT OF ITS EFFECT ON THE FLASH-BACK CHARACTERISTICS OF THREE PURE HYDROCARBONS (TOLUENE, NORMAL HEPTANE AND BENZENE) IN THE PHILLIPS MICROBURNER.

3 2 PREVIOUS STUDIES IN PHILLIPS 2 INCH TURBOJET ENGINE TYPE COMBUSTOR HAD INDICATED THAT SUCH MATERIALS COULD SUBSTANTIALLY INCREASE THE MAXIMUM RATE OF HEAT RELEASE ATTAINABLE, ESPECIALLY WITH LOW PERFORMANCE FUELS SUCH AS THE ISO PARAFFIN TYPE HYDROCARBONS-PARTICULARLY WHEN OPERATING UNDER SEVERE CONDITIONS FOR COMBUSTION (I.E., HIGH AIR FLOW VELOCITY OR LOW COMBUSTION PRESSURE).

4 1 THE ASSUMPTION HAS BEEN MADE IN THIS FUEL EVALUATION THAT THE GREATER THE ALLOWABLE HEAT INPUT RATE FOR A GIVEN VELOCITY, THE GREATER THE DEGREE OF COMBUSTION STABILITY.

4 2 ON THIS BASIS, THE DATA INDICATE THAT ALL THE ADDITIVE MATERIALS TESTED CAUSED AN INCREASE IN STABILITY PERFORMANCE., A FUEL OF RELATIVELY LOW PERFORMANCE SUCH AS TOLUENE BEING BENEFITED TO A GREATER EXTENT THAN A HIGH PERFORMANCE FUEL SUCH AS NORMAL HEPTANE.

4 3 IN GENERAL, ADDITIVE CONCENTRATIONS OF ONE PER CENT BY WEIGHT IN THE SEVERAL PURE HYDROCARBONS WHICH NORMALLY DIFFERED QUITE WIDELY IN PERFORMANCE, PRODUCED UNIFORMLY SUPERIOR COMBUSTION STABILITY CHARACTERISTICS AS MEASURED USING THE PHILLIPS MICROBURNER.

5 0 I. INTRODUCTION

6 1 AT THE REQUEST OF THE NAVY BUREAU OF AERONAUTICS THE JET FUELS GROUP HAS EVALUATED THE EFFECTS OF THE ADDITION OF SMALL AMOUNTS OF DIMETHYLAMINEBORIN =S ON THE COMBUSTION STABILITY PERFORMANCE OF SEVERAL HYDROCARBON FUELS.

7 1 DUE TO THE SMALL QUANTITY OF THIS MATERIAL OBTAINED THE EVALUATION WAS CONDUCTED IN THE PHILLIPS MICROBURNER (MODEL 1A) WHICH IS A SLIGHTLY MODIFIED VERSION OF THE ORIGINAL PHILLIPS MICROBURNER (MODEL 1).

8 0 II. DESCRIPTION OF PHILLIPS MICROBURNER (MODEL 1A)

10 0 III. DESCRIPTION OF TEST APPARATUS

14 0 IV. DESCRIPTION OF TEST FUELS

17 0 V. TEST PROCEDURE

21 0 VI. RESULTS

25 0 VII. DISCUSSION

29 1 PREVIOUS WORK CONDUCTED IN THE PHILLIPS 2 INCH COMBUSTOR (REF. 2) INDICATED THAT SOME ADDITIVES CAUSED A SIGNIFICANT INCREASE IN THE PERFORMANCE OF A LOW RATING FUEL WHILE THESE SAME ADDITIVES DID NOT SUBSTANTIALLY EFFECT THE HIGHER RATING FUELS.

31 3 ALL FOUR ADDITIVES INDICATED THEIR ADDITION TO BE SUBJECT TO THE EFFECT OF DEMINISHING RESULTS UPON FURTHER ADDITION-THAT IS, THEIR EFFECT WAS NOT ESSENTIALLY A BLENDING EFFECT.

33 0 VIII. CONCLUSIONS

34 1 1. THE ADDITION OF DIMETHYLAMINE BORINE IN CONCENTRATIONS OF ONE PER CENT BY WEIGHT TO JET FUEL TYPE HYDROCARBONS RESULTED IN A UNIFORMLY HIGH LEVEL OF COMBUSTION STABILITY PERFORMANCE AS MEASURED BY PHILLIPS MICROBURNER.

35 1 2. THE ADDITION OF RELATIVELY LARGE AMOUNTS OF PROPYLENE OXIDE TO TOLUENE WERE NECESSARY TO PROVIDE SIGNIFICANT IMPROVEMENT IN STABILITY PERFORMANCE AS INDICATED BY INCREASES IN ALLOWABLE HEAT INPUT RATES.

36 1 3. THE ADDITION OF ADDITIVE CONCENTRATIONS (UP TO 1 PER CENT) OF AMYL NITRATE, CUMENE HYDROPEROXIDE, AND DIMETHYLAMINE BORINE ALL RESULTED IN IMPROVED STABILITY PERFORMANCE., THE GREATEST INCREASES WERE SHOWN WHEN BLENDED WITH A FUEL OF POOR PERFORMANCE CHARACTERISTICS-SUCH AS TOLUENE.

37 0 IX. RECOMMENDATIONS

38 1 BASED ON THE EVALUATION OF THE EFFECTS OF ADDITIVES ON THE FLASHBACK LIMITS OF THE ADDITIVE-FUEL BLENDS TESTED IN THE MICROBURNER (MODEL 1A) IT IS RECOMMENDED THAT DIMETHYLAMINE BORINE SHOULD BE FURTHER INVESTIGATED.

38 2 THIS FUTURE WORK SHOULD INCLUDE STUDY OF COMBUSTION STABILITY AND COMBUSTION EFFICIENCY EFFECTS IN THE PHILLIPS 2 INCH COMBUSTOR AND AN INVESTIGATION OF ITS INFLUENCE ON COMBUSTION CLEANLINESS.

Figure 2.10 Example of an abstract produced by Edmundson's system. (Reproduced from The Journal of the Association for Computing Machinery (54).)

demonstrated how an English/Russian phrase data base can be used to develop a technique for obtaining English indexes from untranslated Russian text.

During the third and fourth years of the project, experiments in the compilation of a "sentence dictionary" of syntactic types began and compilation of English syntactic word government tables was undertaken. The hypothesis which the sentence dictionary was used to support, states that if a large group of sentences, as representative of the language as possible, are processed, classified as "indexable" or "nonindexable," and assigned a syntactic structure, then when these structures are sorted, it will be found that like structures have like index classifications. The structures can be ordered into a "dictionary" of sentence types, each classified as indexable or non-indexable. When sentences from a document which is to be abstracted are matched against the sentence dictionary, those sentences which are indexable would be candidates for inclusion in the index or in the abstract. Table 2.3 shows the results of experiments designed to test this hypothesis.

Based on these data, it seemed clear that representing a sentence by part-of-speech strings made too fine a distinction between sentences. The sentences were then structured into phrases, to cause sentences of like phrase structure to be grouped. The phrase structure approach to syntactic patterns gave impetus to development of English syntactic word government tables. These word government tables contained entries which reflected the fact that a word's government pattern is often linked with its semantic meaning, that is, syntactic pattern is a clue

Table 2.3 Statistics of part-of-speech patterns in text (from (59))

Item	Number of Chapters in Data Base			
	3	6	8	9
(1) Number of total patterns represented by more than one sentence	18	25	31	34
(2) Number of total patterns represented by more than one sentence, with a consistent index code	14	15	21	23
(3) Number of total duplicated patterns common to more than one article	3	6	8	12
(4) Number of total duplicated patterns common to more than one article, with a consistent index code	2	3	4	5
(5) Number of one-of-a-kind patterns	1198	2425	2822	3064
(6) Number of total unique patterns	1216	2450	2853	3098
(7) Ratio of the number of one-of-a-kind patterns to number of total unique patterns	0.985	0.989	0.989	0.992

to semantic meaning. Experiments were devised to test the applicability of the phrase to computer-based abstracting. The data obtained from these experiments is shown in Table 2.4 (61). From these experiments it was concluded that both the part-of-speech and the phrase-pattern methods of syntactic classification are inadequate to separate indexable from non-indexable sentences. The experimental results shown in Table 2.4 indicate that there are far too many unique patterns and that the consistency of index codes tends to decrease with the number of unique patterns. Based on these results, it was concluded that indexable and non-indexable sentences cannot be distinguished by structure alone.

During the fifth year of the project, Earl and associates developed a parsing program, initiated some extracting experiments on technical text, and experimented with automatic indexing of a medical book. In the sixth year, the sentence dictionary experiment was concluded, the extracting experiment was completed, a frequency-syntax method of indexing was conceived and tested, and the concept of English syntactic word government was expanded while compilation of the tables continued. During the seventh year, the scope of the parsing program was extended, preparatory to additional indexing experiments using syntax in conjunction with frequency counts or word government criteria. Also, during this period, some studies in describing and abstracting pictorial structures were undertaken. A critical review of the field was

Table 2.4 Comparison of part-of-speech and phrase patterns (from (61)).

	Item	Part-of-Speech Patterns	Phrase Patterns
(1)	Number of total patterns represented by more than one sentence	34	35
(2)	Number of total patterns represented by more than one sentence with a consistent index code	23	15
(3)	Number of total duplicated patterns common to more than one article	12	26
(4)	Number of total duplicated patterns common to more than one article, with a consistent index code	5	11
(5)	Number of one-of-a-kind patterns	3064	3026
(6)	Number of total unique patterns	3098	3061
(7)	Ratio of the number of one-of-a-kind patterns to number of total unique patterns	0.992	0.988

prepared[27] and a series of experiments using human subjects to describe aerial terrain photographs was conducted.

While some interesting experimental data have been obtained by Earl, et al., it appears that they are no closer to a system for computer-based abstracting than when the study began.

2.2.2.7. The Soviet Study

E. F. Skoroxod'ko is carrying on research in automatic abstracting at the Academy of Sciences of the Ukranian SSR in Kiev, USSR (64, 65). His research is based on the assumption that only a method of automatic abstracting which adapts itself to a text can provide good and stable results. This assumption is based on the belief that a given text has individual characteristics and that the optimal selection criteria and abstracting procedure should be determined based on those characteristics.

The individual characteristics of a given text are defined by the form of a semantic network which represents the text. A semantic network is defined to be a graph where the nodes are associated with sentences and the arcs are associated with semantic relations between sentences. Two sentences, A and B, are said to be semantically related if 1) at least one noun occurs in both sentences A and B, or 2) sentence A contains a word "a" and sentence B contains a word "b" where "a" and "b" have been predefined as being semantically related, or 3) when

---

[27] To be published in 1973.

words "a" and "b" in sentences A and B are related with respect to a given text. The graphs of the semantic structures can be identified in terms of four important types which are shown i. Figure 2.11. The selection of an appropriate abstracting procedure is based on the type of semantic structure.

The significance of each sentence is assumed to be directly proportional to the number of sentences which are semantically related to it. Thus, nodes in the graph which have the most incident arcs are defined to be the most significant. The sentence significance also depends upon the amount of change in the semantic network for that document when the node for that sentence is removed from the network. The general significance of a sentence is determined using the following formula:

$$F_i = N_i \, (M - M_i^H)$$

where

F_i - is the functional weight of a sentence in text

N_i - is the number of arcs incident to a node associated with a given sentence (i.e., the number of sentences semantically related to a given sentence)

M_i - is the total number of nodes in a sentence network (i.e., the number of sentences in a text)

M_i^H - is the maximum number of nodes in any connected component of a network after removal of a node associated with a given sentence (i.e., the number of sentences in the

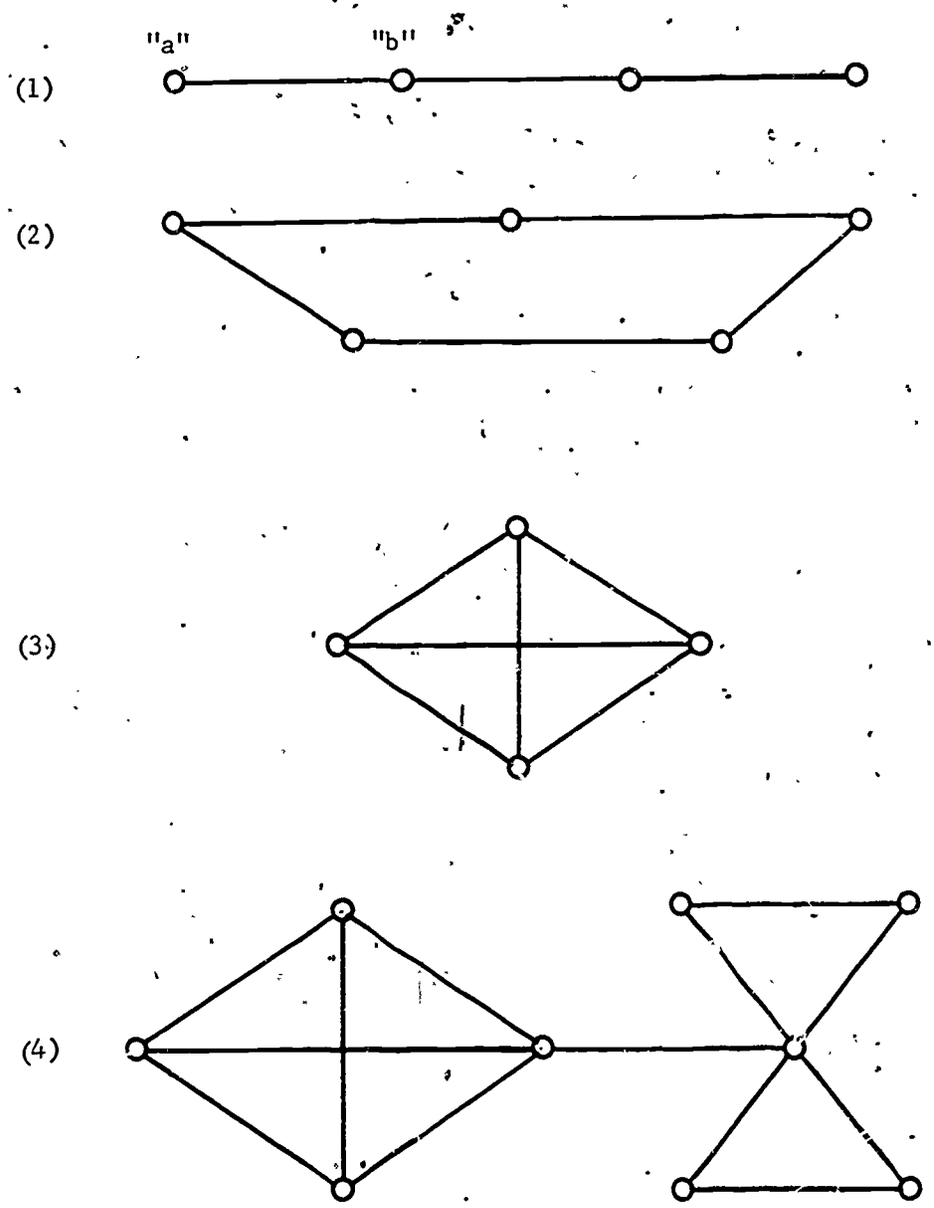


Figure 2.11 Semantic structure types in the Soviet study: (1) Chained, (2) Ringed, (3) Monolithic, and (4) Piecewise.

longest connected fragment of text formed after the removal of given sentence).

Based on this measure, Skoroxod'ko concludes that in the case of chained or ringed structure networks, it is impossible to form an adequate extract from sentences raken from the text since all sentences have approximately the same semantic value. Automatic abstracts can only be generated from texts where semantic relationships can be depicted either as monolithic or piecewise structures.

Skoroxod'ko presents only one of the procedures developed for the adaptable process (64). There are the following seven operations with operations 6 and 7 being optional:

1. The determination of functional weights of all sentences in a text.
2. The compression of a text, i.e., the removal of sentences which have functional weights considerably less than the average functional weight of sentences throughout the text. Such sentences are generally examples, explanations, etc.
3. The segmentation of a text, i.e., its division into segments which are relatively autonomous in semantic and informational aspects. A section begins with a sentence whose linear coefficient is less than a definite critical value.
4. The selection of one or more sentences with maximum functional weight in each segment. A set of such sentences forms an abstract of the text (extract).
5. The determination of functional weights of words in an abstract.

6. The removal of words with minimal functional weights from an abstract.

7. The translation of an abstract into an informational retrieval language (if necessary for information retrieval).

The approach taken by Skoroxod'ko relies heavily on the co-occurrence of words in the text and the matching of words to synonym definitions in the dictionary. The quality of the abstracts would then appear to be greatly dependent on the construction of the dictionary, which is manually produced. Skoroxod'ko's publication (64) does not include any sample abstracts produced by this system nor any mention of evaluation procedures. Thus, although the theory appears to be well developed, it is not possible to ascertain the practical effectiveness of this method.

References

1. B. A. Mathis and J. E. Rush, "Abstracting," in J. Belzer, A. G. Holzman, and A. Kent (Eds.) Encyclopedia of Computer Science and Technology, Marcel Dekker, Inc., New York, New York, 1973.

2. "Abstracting", Encyclopedia Brittanica, Volume 1, Encyclopedia Brittanica, Inc., Chicago, Ill., 1957, 67-68.

3. B. Weil, "Standards for Writing Abstracts", Journal of the American Society for Information Science 21 (5), 351-357 (1970).

4. M. G. Mellon, Chemical Publications: Their Nature and Use, McGraw-Hill Book Co., New York, New York, 1965.

5. F. Bacon, Novum Organum, 1620.

6. D. J. de Solla Prince, Science Since Babylon, Yale University Press, New Haven, Connecticut, 1961.

7. American Chemical Scoiety, Handbook for Authors of Papers in the Journals of the American Chemical Society, American Chemical Society Publication, Washington, D.C., 1967.

8. Chemical Abstracts Service, Directions for Abstractors, The Ohio State University, Columbus, Ohio, 1971 Revision.

9. R. E. Wyllys, "Extracting and Abstracting by Computer" in H. Borko (ed.), Automated Language Processing, John Wiley and Sons, Inc.; New York, New York, 1968.

10. R. E. Maizell, J. F. Smith, and T. E. R. Singer, Abstracting Scientific and Technical Literature, Wiley-Interscience, New York, New York, 1971.

11. H. Borko and S. Chatman, "Criteria for Acceptable Abstracts: A Survey of Abstractors' Instructions", American Documentation 14 (2), 149-160 (1963).

12. J. E. Rush, R. Salvador, and A. Zamora, "Automatic Abstracting and Indexing. II. Production of Indicative Abstracts by Application of Contextual Inference and Syntactic Coherence Criteria", Journal of the American Society for Information Science 22 (4), 260-274 (1971).

13. B. C. Landry and J. E. Rush, "Toward a Theory of Indexing--II", Journal of the American Society for Information Science 21 (5), 358-367 (1970).

14. A System Study of Abstracting and Indexing in the United States. System Development Corporation, Falls Church, Virginia, 1966 (PB 174 249).

15. J. E. Hopcroft and J. D. Ullman, Formal Languages and Their Relation to Automata, Addison-Wesley Publishing Company, Reading, Massachusetts, 1969.

16. E. J. Crane (ed.), The Production of Chemical Abstracts, The American Chemical Society, Washington, D.C., 1959.

17. G. K. Zipf, Human Behavior and the Principle of Least Effort, Addison-Wesley Publishing Company, Cambridge, Massachusetts, 1949.

18. F. B. Libaw, "A New Generalized Model for Information Transfer: A Systems Approach," American Documentation 20 (4), 381-384 (1970).

19. Chemical Abstracts Service, Report on the Fourteenth Chemical Abstracts Service Open Forum, American Chemical Society, Chicago, Ill., 1970.

20. A. Resnick, "Relative Effectiveness of Document Titles and Abstracts for Determining Relevance of Documents", Science 134 (3484), 1004-1006 (1971).

21. G. J. Rath, A. Resnick, and T. R. Savage, "Comparisons of Four Types of Lexical Indicators of Content", American Documentation 12 (2), 126-130 (1961).

22. G. J. Rath, A. Resnick, and T. R. Savage, "The Formation of Abstracts by the Selection of Sentences Part I. Sentence Selection by Men and Machines" American Documentation 12 (2), 139-141 (1961).

23. A. Resnick, "The Formation of Abstracts by the Selection of Sentences Part II. The Reliability of People in Selecting Sentences", American Documentation 12 (2), 141-143 (1961).

24. H. P. Edmundson, V. A. Oswald, and R. E. Wyllys, Automatic Indexing and Abstracting of the Contents of Documents, Planning Research Corporation, Los Angeles, California, Prepared for Rome Air Development Center, Air Research and Development Command, Griffiss Air Force Base, New York, 1959 (AD 231 606).

25. Final Report on the Study for Automatic Abstracting, Thompson Ramo Wooldridge, Inc., Canoga Park, California, 1961 (PB 166 532).

26. H. P. Luhn, "The Automatic Creation of Literature Abstracts", I.B.M. Journal of Research and Development 2 (2), 159-165 (1958).

27. P. B. Baxendale, "Machine-Made Index for Technical Literature--An Experiment", I.B.M. Journal of Research and Development 2 (4), 354-361 (1958).

28. A. Resnick and T. R. Savage, "The Consistency of Human Judgements of Relevance", American Documentation 15 (2), 93-95 (1964).

29. R. Carnap, Logical Foundations of Probability, The University of Chicago Press, Chicago, Illinois, 1950.

30. W. D. Climenson, N. H. Hardwick, and S. N. Jacobson, "Automatic Syntax Analysis in Machine Indexing and Abstracting", American Documentation 12 (2), 178-183 (1961).

31. Special Report on Optical Character Recognition, Auerbach Corporation, Philadelphia, Pennsylvania, 1971.

32. R. G. Casey and G. Nagy, "Advances in Pattern Recognition", Scientific American 224 (4), 56-71 (1971).

33. H. P. Luhn, "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," I.B.M. Journal of Research and Development 1 (4), 309-317 (1957).

34. H. P. Luhn An Experiment in Auto-Abstracting, Auto-Abstract of Area 5 Conference Papers, International Conference on Scientific Information, Washington, D.C., I.B.M. Research Center, Yorktown Heights, New York, 1958.

35. I.B.M. Corporation, Advanced Systems Development Division, ACSI-matic Auto-Abstracting Project, Final Report, Volume 1, Yorktown Heights, New York, 1960.

36. I.B.M. Corporation, Advanced Systems Development Division, ACSI-matic Auto-Abstracting Project, Final Report, Volume 3, Yorktown Heights, New York, 1961.

37. L. B. Doyle, "Semantic Road Maps for Literature Searchers", Journal of the Association for Computing Machinery 8 (4), 553-578 (1961).

38. L. B. Doyle, "Indexing and Abstracting by Association", American Documentation 13 (4), 378-390 (1962).

39. H. Borko, "Indexing and Classification", in H. Borko (ed.), Automated Language Processing, John Wiley and Sons, Inc., New York, New York, 1968.

40. C. L. Bernier and K. F. Heumann, "Correlative Indexes III. Semantic Relations Among Semantemes--The Technical Thesaurus", American Documentation 8 (3), 211-220 (1957).

41. M. R. Quillian, "The Teachable Language Comprehender: A Simulation Program and Theory of Language", Communications of the Association for Computing Machinery 12 (8), 459-476 (1969).

42. H. Borko and M. Bernick, "Automatic Document Classification", Journal of the Association for Computing Machinery 19 (2), 151-162 (1963).

43. M. R. Quillian, Notation for Representing Conceptual Information: An Application to Semantics and Mechanical English Paraphrasing, System Development Corporation, Santa Monica, California, 1963.

44. M. R. Quillian, "Word Concepts: A Theory and Simulation of Some Basic Semantic Capabilities", Behavioral Science, 12 (5), 410-430 (1967).

45. M. R. Quillian, "Semantic Memory", in M. Minsky (ed.), Semantic Information Processing, The M.I.T. Press, Cambridge, Mass., 1968.

46. A. Avramescu, "Probabilistic Criteria for the Objective Design of Descriptor Languages", Journal of the American Society for Information Science 22 (2), 85-95 (1971).

47. R. Fugmann, H. Nickelsen, I. Nickelsen, and J. H. Winter, "TOSAR--A Topological Model for the Representation of Synthetic and Analytic Relations of Concepts", Angewandte Chemie International Edition in English 9 (8), 589-595 (1970).

48. Appendix "D"--Final Report on the Study for Automatic Abstracting, Thompson Ramo Wooldridge, Inc., Canoga Park, California, 1961 (PB 166 533).

49. H. P. Edmundson and R. E. Wyllys, "Automatic Abstracting and Indexing: Survey and Recommendations", Communications of the Association for Computing Machinery 4 (5), 226-235 (1961).

50. H. P. Edmundson, Automatic Abstracting, TRW Computer Division, Thompson Ramo Wooldridge Inc., Canoga Park, California, 1963 (AD 406 155).

51. H. P. Edmundson, "An Experiment in Abstracting Russian Text by Digital Computer," in H. P. Luhn, (ed.), Automation and Scientific Communication, American Documentation Institute, 26th Annual Meeting, Chicago, Ill., October 1963.

52. R. E. Wyllys, Research in Techniques for Improving Automatic Abstracting Procedures, System Development Corporation, Santa Monica, California, 1963 (AD 404 105).

53. H. P. Edmundson, "Problems in Automatic Abstracting", Communications of the Association for Computing Machinery 7 (4), 259-263 (1964).

54. H. P. Edmundson, "New Methods in Automatic Extracting", Journal of the Association for Computing Machinery 16 (2), 264-285 (1969).

55. J. L. Dolby, L. L. Earl, and H. L. Resnikoff, The Application of English-Word Morphology to Automatic Indexing and Abstracting, Lockheed Missiles and Space Co., Palo Alto, California, 1965 (AD 615 424).

56. B. D. Rudin, Automatic Indexing and Abstracting Part I, Lockheed Missiles and Space Co., Palo Alto, California, 1966 (AD 631 241).

57. B. D. Rudin, Automatic Indexing and Abstracting Part II. English Indexing of Russian Technical Text, Lockheed Missiles and Space Co., Palo Alto, California, 1966 (AD 631 242).

58. L. L. Earl, Annual Report: Automatic Indexing and Abstracting, Lockheed Missiles and Space Co., Palo Alto, California, 1967 (AD 6[illegible] 057).

59. L. L. Earl and H. R. Robinson, Automatic Informative Abstracting and Extracting, Lockheed Missiles and Space Co., Palo Alto, California, 1968 (AD 667 473).

60. L. L. Earl and H. R. Robinson, Automatic Informative Abstracting and Extracting, Lockheed Missiles and Space Co., Palo Alto, California, 1969 (AD 696 653).

61. L. L. Earl and H. R. Robinson, Automatic Informative Abstracting and Extracting, Lockheed Missiles and Space Co., Palo Alto, California, 1970 (AD 867 656).

62. L. L. Earl, "Experiments in Automatic Extracting and Indexing", Information Storage and Retrieval 6 (4), 313-334 (1970).

63. L. L. Earl, D. Firschein, and M. A. Fischler, Annual Report: Automatic Informative Abstracting and Extracting, Lockheed Missiles and Space Co., Palo Alto, California, 1971 (AD 721 066).

64. E. F. Skoroxod'ko, "Information Retrieval System of the Ukranian Institute of Cybernetics", in International Congress on Scientific Information, Moscow, 1968.

65. E. F. Skoroxod'ko, "Adaptive Method of Automatic Abstracting and Indexing", IFIP Congress 71, Ljubljana, Yugoslavia, 1971, Booklet TA-6, 133-137.

CHAPTER III. DESIGN OF THE ABSTRACTING SYSTEM

The automatic abstracting system which has been used as the basis for this research has been described in detail elsewhere (1, 2, 3). This chapter provides a brief description of the system as it was originally developed by J. E. Rush, R. Salvador, and A. Zamora. The system has been named ADAM, for Automatic Document Abstracting Method.

1. Philosophy Underlying the Abstracting System

The computer-based abstracting system which has been developed consists of two important components, a dictionary, called the Word Control List (WCL), and a set of rules for implementing certain functions specified for each WCL entry. This combination of rules and dictionary has been designed and implemented to accomplish the production of abstracts which are characterized as follows.

a. Their size is approximately 10% of the original document, and the use of arbitrary cut-off criteria is avoided.

b. They use the same technical terminology as in the document.

c. Except for actual results, they contain no numbers or cardinal expressions.

d. Unconventional or rare characters or abbreviations are excluded.

e. Preliminary remarks, equations, footnotes, references, quotations, tables, charts, figures, graphs, descriptive

cataloging data and the like are not included.

f. Negative results, unless they are the sole results, are excluded.

g. They do not contain methodologies of data gathering, measurements, preparation of samples, etc., unless these are the purpose of the work.

h. No examples, explanations, speculative statements, opinions or comparisons are included.

These are mainly things to be excluded from the abstract. On the positive side, it is desirable that the abstract include:

i. Objectives of the work.

j. Methods used in the work (if they are the main purpose of the investigation).

k. Results and conclusions.

To automatically produce abstracts with these characteristics, it is necessary to identify and eliminate certain sentences of the document. It is also necessary to identify and select a few sentences for the abstract, and to retain, by default, certain sentences for the abstract. These three methods of sentence handling are discussed in Section 1.3.

For efficiency, a language processing program should consider the largest independent item in its data base as its basic unit. Thus, in automatic abstracting the basic unit is an original article. Any approach which considers paragraphs or sentences as basic units is inadequate because there is an interdependence between these elements and the remainder of an article. A program which operates on inter-

dependent units bears the burden of carrying data from one unit to the next. An automatic language processing program must also be able to identify and manipulate the linguistic elements of its basic data unit, whether these elements be words, phrases, clauses or sentences.

2. Data Input

Data input involves two basic processes: the pre-editing of the original document and preparation of the data in a format acceptable to the abstracting system. These processes should involve little or no human intervention.

Pre-editing means pre-processing editing. The text to be abstracted is edited in some way prior to its being processed by the abstracting programs. In the work of Edmundson, et al. (4), pre-editing involved the insertion of special markers (flags) into the input text to delimit sentences, paragraphs, section headings, etc. This type of pre-editing has been considered necessary because of the ambiguous usage of periods, commas, etc. ADAM eschews this type of pre-editing altogether in favor of automatic recognition of phrases, clauses, sentences and other elements of the original document. On the other hand, when manual data input is employed, the keyborad operator is provided with instructions to omit figures, tables and other similar graphic material which will not, in any event, find its way into the abstract. Thus ADAM employs minimal pre-editing of the input text.

Similarly, ADAM requires no special formatting of the input text. The system accepts the text as a continuous string and performs all partitioning (into words, sentences, etc.) of the input text automatically. It can be made to accept the data in any particular code (ASCII, BCD, EBCDIC, etc.) with only minor modification of the system.

3. Methods of Sentence Selection and Rejection

In most research on computer-based abstracting, it has been considered necessary to analyze the conditions under which various methods of sentence selection are successful, in order to develop criteria for selecting sentences to form an abstract (5). But clearly an abstract can also be produced by rejecting sentences of the original document which are irrelevant to the abstract. Rush, Salvador and Zamora found that methods for rejecting sentences were found to be more fruitful than selection methods in most cases (1). It is upon this idea that ADAM is built. In the following sections, methods of sentence selection and rejection are discussed, including contextual inference, intersentence reference, frequency criteria, and coherence considerations.

3.1. Sentence Elimination

The exclusion of sentences from the abstract involves the detection of words or strings of words which identify sentences giving historical data, results of previous work, examples, explanation, speculative material and so on. Analysis of documents from a number of scientific disciplines showed that the set of word strings needed

to accomplish this task need not be large, perhaps a few hundred word strings serving to eliminate up to 90% of the sentences of a document. Such word strings are incorporated in a dictionary called the Word Control List. Sentence elimination is not, however, carried out blindly. An important aspect of sentence elimination is the location within the sentence of the "offending" word or phrase. This aspcet is considered under "location criteria."

3.2. Location Criteria

Location criteria for sentence rejection (or selection) are based on the physical arrangement of the linguistic elements of an article. This arrangement can be described in terms of the location of a sentence with respect to the limits of its containing document, or in terms of the location of phrases, clauses, or words with respect to the limits of a sentence.

The first of these arrangements (sentence location) is governed by the style of the author or the editor, with general writing guides providing advice about the placement of sentences within an article. Sinct it is not possible to dictate in the matter of style, the location of a sentence does not convey an unequivocal criterion for sentence selection or rejection.

The second location type is really a sentence description. The location of phrases and words within the sentence is subject to grammatical rules to which authors and editors adhere. Even a partial analysis of a sentence yields its basic structure, since the number

of basic sentence frameworks is small.

Punctuation marks are important data for use with location criteria. A sentence terminated by a question mark can be rejected out of hand, and sentences located in close physical proximity to it can also usually be rejected because they are related directly to the question. Commas are used to delimit clauses and lists (series of items), and to separate digits in a number. Periods not only delimit sentences, they also appear in numbers, ellipses and abbreviations. These usages of punctuation must be differentiated in order to avoid error. The way in which location criteria are employed in ADAM will be made clearer below.

3.3. The "Cue" Method of Sentence Rejection and Selection

ADAM uses what Edmundson (6) called the "Cue" method almost exclusively. Cue words are words or strings of words which are, in general, unequivocal clues to such things as opinion and subjectivity, as well as to some positive notions. In this system, Cue words are contained in a dictionary called the Word Control List (WCL) (see Section 4), together with codes which indicate their function within a sentence, or within a particular location in the sentence.

The Cue method provides a powerful approach to sentence selection or rejection. The method depends on the fact that it is possible to decide what should or should not be included in an abstract, based upon the presence in the original article of particular words or combinations of words. For example, words which are known to be used

in sentences which state the purpose of a paper serve to indicate that such sentences should be selected for the abstract. "Our work", "this paper", and "present research" are expressions which meet this criterion, but so does "this theme paper". It is necessary, then, to effect partial matches between WCL entries and words of a sentence to allow for varied input while maintaining a manageable WCL. A partial match occurs when one or more words intervene between any two words of a cue expression.

Opinions, references to figures, and other items which should not be included in an abstract can be identified by cue words such as "obvious", "believe", "Fig.", "Figure 1", "Table IV", etc. The weight of a cue word may also depend on its position in a sentence. A sentence starting with "A" or "Some" is more likely to present detailed descriptions than a sentence which contains either of these words in a more central location of the sentence. This is because these words have a strong quantitative function when they appear at the beginning of a sentence. Similarly, sentences which begin with participles are usually conditional in nature, indicating assumptions or conjectures.

Thus, the Cue method is based upon the identification within a sentence, and also within a particular location in the sentence, of words or word strings found in a dictionary. It is important, therefore, that the dictionary be kept small and constant. Neither a large dictionary nor a rapidly changing one will permit the development of a viable, operational automatic abstracting system. These considerations weighed heavily in favor of the decision to employ rejection criteria

almost exclusively. Cue words which indicate that a sentence contains nothing of import are few in number, and of stable and unambiguous usage. By contrast, the cue words which indicate important notions are many in number and are of variable and ambiguous usage. By using the rejection approach to abstract production, the Word Control List has been made to contain fewer than 700 entries.

3.4. Intersentence Reference

Intersentence references give much information about the logical relationships within the text material, but they require involved treatment if a coherent abstract is to be produced. If more than one clause exists in a sentence, then the first clause is indispensable to the meaning of the sentence. The first clause will usually contain intersentence references if there are any. Words in the second and subsequent clauses which require antecedents usually refer to the first clause. Some cue words that indicate intersentence references are "these", "they", "it", and "above". When these words have multiple uses, additional criteria are required to determine if there is intersentence reference.

A special case of intersentence reference is that between the title and the sentences of a document. The Title method has as a premise that the author (generally) describes in as few words as possible the essence of his paper; it can be assumed, then, that the words of the title are well chosen and of high significance.

In ADAM, the words of each sentence are matched against the potential-information-carrying words of the title; if any words match, the sentence becomes a candidate for selection. However, the possibility that the substantive words of the title may appear frequently, means that additional criteria should be used before any sentence containing words which also occur in the title is accepted for inclusion in the abstract.

3.5. Frequency Criteria

A simple yet effective way of introducing frequency criteria is as follows: if any cue expression exceeds a given frequency threshold, then its value should be reduced. This means that if the cue expression has a positive weight it should become less positive, and if it has a negative weight it should become less negative. With these guidelines it should be possible to successfully produce abstracts of papers in which cue words are used in unusual ways. The thresholds at which the weight transitions should take place need to be determined, but statistical data are needed only for the cue expressions contained in the dictionary. In ADAM a module has been incorporated, which is optionally executable, which decreases the strength of both negatively and postitively weighted WCL entries (i.e., makes the entry less influential) when the WCL entry is found more frequently than desired in the text.

3.6. Coherence Considerations

Regardless of the portion of the original document that finds its way into the abstract, that abstract should be as coherent as possible both logically and linguistically. Thus the progression of ideas presented in the abstract should flow smoothly and each sentence of the abstract should be linguistically well-formed. This latter criterion requires some analysis of the sentences selected for the abstract before the final set can be determined. At present, ADAM merely checks each sentence for the presence of a verb and rejects that "sentence" if no verb is found.

4. The Word Control List (WCL)

The WCL consists of an alphabetically ordered set of words and phrases, which are referred to collectively as word strings, and one or two associated codes. The entries in the WCL are treated as functions and each has two arguments: a semantic weight and a syntactic value. Each function returns a value which indicates whether the sentence is a candidate for retention or deletion. In general, a WCL entry is represented as

WORD STRING*([semantic weight]*[syntactic value])

where WORD STRING is a string of alphanumeric characters and blanks, the *'s are delimeters and semantic weight and syntactic value are one character fields. The parentheses and brackets in the above expression are used to indicate that the presence of the enclosed items are

required or optional, respectively; they do not appear in the WCL. As will be seen later, entries in the WCL may be varied as desired without necessitating any change in the programs of the system.

4.1. Hierarchical Rules for Implementing the Functions in the WCL

When one attempts to determine whether a sentence of the document is a member of the abstract or not, using a two-valued membership criterion, it is often found that a sentence is both an element, and yet not an element, of the set of sentences constituting the abstract because it contains word strings of both positive and negative semantic weight. Two alternate solutions of this predicament are available: 1) impose an ordering on the several semantic weights, or 2) determine a degree of membership of the sentence in the abstract (7). The former alternative has been chosen for implementation, but it would be interesting to compare abstracts obtained using each of the methods. The effect of this choice is to both simplify and speed up processing.

The hierarchy of rules for implementing the semantic weights from the WCL is shown in Table 3.1. Considerable flexibility is obtained by incorporating the rules in the program and supplying the semantic weights externally. The rules can be altered without the necessity for changing the WCL, and the WCL can be altered independently of the rules.

4.2. Syntactic Values and Their Use

Since the implementation of the semantic weights discussed in the previous section requires some syntactic information, a partial

Table 3.1. Semantic attributes for WCL entries

Semantic Code[a]	Description
I	Used for very positive terms; those which almost unequivocally indicate something of importance (e.g., our work)
A	Assigned to very negative terms; terms which do not belong in an abstract (e.g., obvious, previously)
K	Assigned to terms which are related to items of positive data content (e.g., important)
B	Parenthetical expressions, terms of low data content, or terms which are associated with items of low data content (e.g., however)
E	Used for intensifiers and determiners (e.g., many, more)
L	Introductory qualifiers (e.g., once, a)
C	Used for words which require an antecedent (e.g., this, these)
H	Terms which introduce a modifying phrase or clause (e.g., whose)
F	Null (assigned to abbreviations)
G	Assigned by the program to indicate intersentence relationships or relation of sentence to title
J	Continuation of a semantic code assigned previously
D	Delete a word (can be used with any arbitrary WCL entry)

[a] Listed in descending order of priority.

syntactic analysis of each sentence is performed. This analysis is carried out through use of the syntactic values associated with entries in the WCL, in conjunction with procedures implemented within the program. One of ten possible syntactic values may be associated with an entry in the WCL. These values are shown in Table 3.2, together with their meanings.

Principal use of the syntactic values is made in an analysis of the commas in a sentence. For effectively utilizing contextual inferences, as discussed earlier, three types of commas are distinguished (numerical commas (e.g., in 12,732) are masked at input): 1) commas which separate phrases, here called "real" commas; 2) commas which separate the elements of a series, called "serial" commas; 3) commas which set off dependent clauses, called "parenthetical" commas.

Parenthetical commas cause the phrase or clause they isolate to be deleted. Serial commas are masked to prevent their being confused with real commas in the later program steps, but are unmasked at output. Real commas delimit phrases or independent clauses and this information is used in implementing some semantic rules.

5. Summary of the Operation of the Abstracting System

The abstracts that are produced by ADAM can be characterized by the following six observations.

1. The terminology is the same as the parent document.
2. The style is the same as the parent document.
3. No figures, graphs, footnotes, or examples are included.

Table 3.2. Syntactic values for WCL entries

Syntactic Code	Description
A	Article
C	Conjunction
D	Delete the word
F	Null word
J	Continuation of a previous syntactic value
N	Pronoun
P	Preposition
O	Exclusively assigned to OF
Q	Exclusively assigned to TO
R	Exclusively assigned to AS
V	Verb
W	Auxiliary verb
X	Exclusively assigned to IS, ARE, WAS, and WERE
Z	Negatives

4. The size is about 10% of the parent document, on the average.
5. Negative results, unless they are the only results, are excluded.
6. Objectives, results, and conclusions are included in the abstract.

The system has been designed to accept documents of different subject areas. The abstracts that are produced by ADAM vary in quality, but this variation appears to be due to the differences in quality of the parent document rather than to differences in subject matter. Concerning differences in subject matter, it is possible to produce abstracts which reflect a particular point of view or subject area through deliberate variation of the Word Control List. Such tailor-made abstracts could be produced by varying the weights of existing WCL entries and by adding entries which reflect the desired viewpoint.

The length of the abstract is not an input parameter to ADAM or a criterion in the selection process. The lengths of abstracts tend to fall within a range of 0 - 35% of the length of the parent document. A long abstract usually reflects that the parent document had many clearly-stated ideas. The length of the abstracts can be modified indirectly by altering the contents of the WCL. For example, to reduce the average length of the abstracts, the WCL entries should be constructed so that the selection criteria are more stringent and the rejection criteria are more relaxed.

The abstracts produced by ADAM seem to be quite good and the efficiency and effectiveness of the selection criteria are very encouraging. The next step in the design of a computer-based abstracting system is to develop a method for evaluation of the quality of the abstracts which will indicate the direction for further improvements in the system.

References

1. J. E. Rush, R. Salvador, and A. Zamora, "Automatic Abstracting and Indexing. II. Production of Indicative Abstracts by Application of Contextual Inference and Syntactic Coherence Criteria," Journal of the American Society for Information Science 22 (4), 358-367 (1971).

2. R. Salvador, Automatic Abstracting and Indexing. The Computer and Information Science Research Center, The Ohio State University, Columbus, Ohio, 1969 (OSU-CISRC-TR-69-15).

3. A. Zamora, System Design Considerations for Automatic Abstracting, Master's Thesis, The Ohio State University, August 1969.

4. H. P. Edmundson, J. Brewer, L. Ertel and S. Smith, Progress Report No. 3 Computer-Aided Research in Machine Translation--Experiment in Automatic Abstracting of Russian, National Science Foundation Contract NSF-C233, June 20, 1963, TRW Computer Division, Thompson Ramo Wooldridge, Inc., Canoga Park, California.

5. H. P. Edmundson, "New Methods in Automatic Extracting," Journal of the Association for Computing Machinery 16 (2), 264-285 (1969).

6. Final Report on Study of Automatic Abstracting, Thompson Ramo Wooldridge, Inc., Canoga Park, California, 1961 (PB 166 432).

7. L. A. Zadeh, "Fuzzy Sets," Information and Control 8 (4), 338-353 (1965).

CHAPTER IV. THE EVALUATION OF THE QUALITY OF ABSTRACTS

1. The Need for an Evaluation Procedure

In the design of any information processing system, there is a need for a method of evaluating that system. In the preceding chapter I described an information processing system which inputs documents and outputs indicative abstracts. Although I have claimed that high quality abstracts result from this system, such a claim must be backed by an objective evaluation of their quality. The evaluation method must be applicable not only to abstracts which result from this system, but to those produced by other abstracting systems as well.

There are two basic reasons for the development of an objective evaluation of the quality of abstracts. First, it is necessary to determine whether computer-produced abstracts are of sufficiently high quality to be used as a substitute for manually produced abstracts. Second, if computer-produced abstracts do not equal or surpass manual abstracts in quality, then it is desirable to learn in what areas the abstracting systems can be improved in order to produce high quality abstracts. Although several methods of evaluation have been proposed, these methods have assumed that manual abstracts should serve as the standard of comparison and have not provided a set of objective criteria which can be used to evaluate and improve computer-based abstracting systems.

It is the aim of this chapter to present a criterion for the objective evaluation of abstracts. This chapter presents 1) a review and analysis of existing methods of evaluation, 2) the requirements for an evaluation criterion, 3) the relation of data and information to abstract evaluation, 4) the definition of the data content criterion, and 5) examples of the application of this criterion.

2. Existing Methods of Abstract Evaluation

Many methods for evaluating the quality of abstracts have been considered and many are currently in use. Perhaps the most widely used source of data for evaluation is user response. Secondary services must produce abstracts which are acceptable to their users if they intend to sell abstract journals. The ultimate test of user acceptance is in the market place, and although the number of subscriptions sold reflects more than just the quality of the abstracts, it is an extremely important factor in the successful publication of abstract journals. Nevertheless, it is highly desirable to have evaluation criteria that can be applied before one enters the market place and prior to the development of a crisis situation. Hence, when I use the term evaluation, I mean the application of tests to the abstracts produced by the system that will indicate whether they satisfy the previously established criteria for acceptability. Of course the market place test will eventually be applied, but with a greater degree of assurance that it will be met with success.

Since manually-produced abstracts are frequently used as standards with which to compare computer-produced abstracts, it is worthwhile to consider the evaluation techniques used within the manual production of abstracts. Most abstracting services use editors to review abstracts for any errors or deficiencies. This process is usually both an evaluation and an improvement of the abstracts. The process of editing is designed to help insure the production of quality abstracts and to provide for a consistency of style in the abstract journals. Manual editing of computer-produced abstracts could be used, but there is a need for an evaluation method which can indicate algorithmic changes in the abstracting system which will produce improved abstracts.

The real problem is 1) the need for an evaluation technique which does not require a standard model for comparisons and 2) the need for a constructive evaluation technique: one that says "fair" or "foul" and in the latter case says why. In other words, the technique must say where deficiencies exist and indicate how to correct them. User evaluations and editor's comments, although helpful, are inadequate for this task for they tend to relate to specific articles and abstracts and are difficult to generalize into rules with broad applicability.

Various researchers have attempted to define evaluation techniques which will result in methods for improving computer-based abstracting systems. Edmundson lists the following five methods of evaluation which the Thompson Ramo Wooldridge (TRW) group considered in their study of automatic abstracting (1):

1. Intuitive value judgement

2. Creation of "ideal" abstracts to serve as standards of comparison.
3. Construction of test questions about the document to be answered from the abstract by a sample population (evaluation of the summary function of an abstract).
4. Retrievability of the document via the abstract (evaluation of the retrieval function of an abstract).
5. Statistical correlations (applicable only to extract-type abstracts).

One or more of these five methods has been used, in some form, by all of the other researchers in automatic abstracting.

All researchers use the intuitive evaluation to some degree when they document the potential values of their system and some seem to use it exclusively. Rush, Salvador and Zamora claim that ADAM produces high quality abstracts, a judgement based on their experience with the production of abstracts at Chemical Abstracts Service (2). Intuitive value judgements, although widely used, can never serve in the place of a uniform, objective criterion. DeLucia constructed guidelines to be used by the abstractor while writing his abstract. These guidelines then served as a standard for evaluation of the final abstract (3). Edmundson created target extracts and did statistical correlations between the computer produced abstract and the target extract (1, 4). Edmundson also used some of the other methods listed above. Payne, Altman, and Munger performed evaluation experiments by asking two groups of college students to answer test questions about a document

after having been presented with either the document or its abstract (5, 6). Although various methods of evaluation have been used, no one method has emerged as an acceptable standard. This is probably because none of the methods has a firm theoretical rationale or provides a measure for the primary factors which contribute to the quality of an abstract.

3. Analysis of Previous Evaluation Measures for Specific Uses of Abstracts

Previous evaluation efforts have been aimed at evaluating the effectiveness of abstracts that have been designed to serve specific functions. The adequacy of these evaluation techniques can be examined in light of the intended use of the abstract. Abstracts serve as accurate, abbreviated representations of documents (7). Within this general purpose, there are four identifiable areas where abstracts are used for specific functions. Abstracts may serve first, as an alerting tool. The user scans the abstract to determine if the document is relevant to his interests. After perusing the abstract, the user should be able to decide whether to read the document or not. The abstract which is to be used as an alerting tool might appear on the first page of the article it abstracts, as the output of an information storage and retrieval system, or in a publication containing abstracts. Second, abstracts may serve as a retrospective search tool. A researcher can scan a set of abstracts to locate relevant documents and to retrieve specific data. The classification and indexing of abstracts provides access to the appropriate abstracts for review. Third,

abstracts may be used by indexers as the source of index terms. Using the abstract can save the indexer time because he need not scan the entire document. Algorithms to produce indexes by computer have also been designed to be applied to abstracts. Fourth, abstracts may serve as the data base in an information storage and retrieval system. The abstracts can be searched for terms that match a user's query. Also, the abstracts may be presented to the user to allow him to determine if that set of documents is relevant. Use of abstracts in an automated information storage and retrieval system results in a significant decrease in storage and search requirements over full text storage and search. Poor quality abstracts can result in degradation of system performance. All of these applications require that the abstracts provide an accurate representation of the content of the document they represent. Let us consider the evaluation of abstracts which will serve in each of these four functions.

3.1. The Abstract as an Alerting Tool

Abstracts can be used to alert a scientist to the existence of a document and to provide him with an indication of its content. The abstract should appear at about the same time as publication of the complete document to be particularly effective. Because abstracts which serve as alerting tools must be published with a minimum of delay, it would be beneficial to produce abstracts by computer from the machine-readable form of the document available from original page composition. Use of computer-produced extracts or abstracts

would provide a reduction in cost and time spent in producing abstracts.

One method, used by Edmundson (1), of evaluating extracts which are to be used as an alerting tool is to compare the test extract with an "ideal" target extract for the document. The differences between the test extract and the target indicate areas of deficiency of the test extract. The best possible test extract would include all of the sentences of the target and exclude all extraneous sentences. The comparison between the test and the target is made on a sentence by sentence basis.

The test extract may be compared to the target extract by means of statistical correlations. The degree of similarity between the test and the target extract is based on a statistical correlation function. Edmundson presents a coefficient of similarity between extracts that is defined in terms of the number of sentences selected in common, the number of sentences selected for the test extract, the number of sentences selected for the target extract, and the total number of sentences in the original document (1). This method assigns a numerical value to each test extract based on its similarity to the target. This method provides an objective ranking of all the test extracts based on the target extract as the standard for comparison.

There is one inherent problem in this method of evaluation, that is, how to find the ideal extract to use as a target. There is always the possibility that several extracts adequately describe the document. The target extract must be compared to all other possible extracts by means of an objective evaluation criterion and must be found to be superior

before it can serve as the ideal. Evaluating the target extract can become as great a problem as evaluating the text extract. Another deficiency of this method is that it can only be applied to extracts. It is inapplicable to abstracts because modification of the sentences of the original document make it impossible to have any exact matches between sentences of the test extract and the target.

3.2. The Abstract as a Retrospective Search Tool

Many secondary services publish journals which contain abstracts (as well as indexes to the abstracts). These services attempt to provide comprehensive coverage of a given subject area. Users search through these volumes to find references to all previous literature which is relevant to their current work. They are often looking for the development of certain trends or the introduction of specific data. This use of abstracts has been modeled by some researchers by setting up a test situation with controlled variables.

Abstracts have been used as the source of answers for test questions in several experiments (5, 6, 8, 9). These experiments are designed to test the amount of data contained in the abstract as compared to the document and to test the ability of a group of subjects to answer questions based on this data. These experiments are usually constructed to reflect a hypothetical user's experience in getting information from a document, but with controls on many of the variable factors. All of the subjects are confronted with the same decision situation and the same data on which to rely. The experiments attempt

to measure the variables of comprehension and the ability to relate data.

This type of experiment was used by Rath, Resnick, and Savage to test the usefulness of two types of abstracts as compared with the usefulness of the complete text or of just the title (8). The subjects were tested to ascertain whether they were able 1) to determine whether a document was relevant for a specific purpose and 2) to find out some data from the document without having to read it in its original form. From the results of these experiments, they concluded that "there was no major difference between the Text and Abstract groups in their ability to pick appropriate documents, but the Text group obtained a significantly higher score on the examination" (8). These results were interpreted by the authors as a function of subject population, test criterion, and document population.

This type of experiment should be carried out with fewer uncontrolled variables to be more effective. The best abstract in this situation would indicate the relevance of the document to a specific purpose and would provide all of the answers to the test questions in the shortest length. Since these factors are determined by questions on an examination, the user needs to know enough data to answer each of the questions. The abstracts could be evaluated in the same manner that the students' examination papers were graded, based on the number of questions where a suitable answer was provided. Subjects present additional variables because of their ability to read, understand, and remember all items of the abstract. The results of the experiment

are influenced by more than just the quality of the abstracts. Experiments which test the quality of abstracts based on the ability of subjects to answer questions on the examination rely on the assumption that the questions presented indicate the most significant part of each article. This assumption may be true for some decision makers, but it will not be true for all decision makers. This method of evaluation requires a large expenditure of effort to construct the examinations and to conduct the experiments, but the results of this evaluation do not appear to justify this amount of effort.

3.3. The Abstract as a Source of Index Entries

Abstracts can be used by an indexer instead of the original document as a source of index terms (10). When an abstract is to be used in this manner, it must serve as an accurate representation of the document. It must be sufficiently complete to serve in place of the document and it must contain all of the significant index terms that could have been derived from the original document. The best abstract for this application would be the one that contained the most index terms in the least amount of length. For this type of application a telegraphic abstract might prove to be more suitable than an abstract in paragraph style. However, the abstract which contains the most index entries might not be the best abstract for other purposes.

3.4. The Abstract as a Data Base for Information Storage and Retrieval Systems

Abstracts are used in some information storage and retrieval systems

to stand as condensed representations of the documents that are available through the system. These abstracts are used to indicate that a relevant document is available to match a user's query. A user requests certain information by formulating the topics which interest him into a query. The form of the query differs with the system, but all queries must express the essential items involved in the user's request. The query is then matched against the abstract to determine if there is any similarity. The abstracts which match the query are usually presented to the user along with the reference to the complete document.

The abstract must serve two distinct functions here. First, it must provide adequate data to the system to allow retrieval of the document it represents if and only if that document is relevant, according to the system's criterion, to the user's query. Second, the abstract must provide adequate data so the user can determine if the document is indeed relevant to his need. In this application it is important to keep the abstracts as short as possible. Additional length would add to the cost of storage and to the time for search and would require additional time for the reader to scan the abstracts.

When the abstract serves a retrieval function, its efficiency can be measured by comparing it with the original document. For example, consider the case presented by Edmundson (1). For a given system suppose there exist two documents A and B. In response to a given query both A and B are retrieved. The user who submitted the query informs the system that only A is really relevant to his request. If

we replace document A by its abstract A' and document B by its abstract B', then there are four possible results:

1. The system retrieves both A' and B'. This represents the same system performance.
2. The system retrieves A' and not B'. This represents a system improvement. The relevant document has been retrieved but noise has been eliminated.
3. The system retrieves neither A' nor B'. This represents a null retrieval situation.
4. The system retrieves B' but not A'. This situation is the worst case. There is no relevant data, only noise.

If A' and B' are extracts, then a document that was not retrieved in the full text search cannot be retrieved in the extract search. If A' and B' are abstracts, there is a possibility of additional, extraneous data being included in the abstract which was not in the document.

The user's query may be compared to the abstract for evaluation. The abstract should include all of the terms of the query that are also in the original document without including any extraneous terms. For a large set of queries the best set of abstracts could be determined by finding those abstracts which provided the best performance for the greatest number of queries.

Once the abstract is retrieved as being relevant to the user's query, the user must use the abstract to determine if the document is indeed relevant. This determination is based on his ability to read

the abstract and to make a decision as to whether he should or should not consult the original. This function of an abstract is similar to other occasions where it serves as an alerting tool, as discussed in Section 3.1.

The methods for abstract evaluation that have been developed do not provide an objective method of evaluating abstracts. All are defined according to specific applications of abstracts and all rely, at some point, on a subjective determination of quality. A general evaluation criterion is needed which can be used for several different purposes and which reflects a user's experience, but does not rely on his opinions. This criterion is needed particularly for the design and improvement of computer-based abstracting systems. In the following sections, I will attempt to define such an evaluation criterion based on a strong theoretical foundation.

4. Desiderata of an Evaluation Criterion

A person records the results of his research and publishes this data in order to communicate his findings to other individuals. He records and communicates the data because it may possibly be of value to other individuals. Information can be derived from the data by an individual if, in some decision making situation, the data is valuable in his choice among courses of action.

The purpose of an abstract of such a research record is to convey some of the data contained in the document to a receiver. The word "some" is used since, by definition, an abstract is an abbreviation of

its parent document and therefore would be unlikely to contain all of the data in the parent document. An abstract is produced because it is expected to be of value to someone. A user may consult an abstract to determine if the document it represents contains some items of data that will be valuable to him in making his decision. The abstract would be of value to him if it indicated either that the document does or does not contain the data. A user may also consult the abstract to locate a specific item of data. The abstract would then be of value to him if the item was contained in the abstract.

Thus, the evaluation of the abstract should determine if the abstract will satisfy the decision maker's need for data. Furthermore, since the abstract will likely be used by many different decision makers, evaluation of the abstract should account for this possibility.

5. The Importance of Data

The abstract presents data to the user. Data results from measurement or observation. Meausrement need not be thought of as requiring determined action by a person. Most measurement is certainly very casually done, perhaps involuntarily done, in fact. For example, if a person inserts a thermometer into some water and observes that the column of mercury of the thermometer corresponds to the mark at 100° C, then he has made a measurement of the temperature of the water. If he puts his finger in the water and declared "This water is too hot to take a bath in!", then he has once again measured the temperature for bathing. Thus, the presence of numeric values is not necessary to

indicate that a measurement has been made and an item of data is present.

In an abstract the sentence, "The ABC process yields a product of 20 percent greater durability as judged by standard test #1234.", provides the reader with data as to the results of the measurement of the ABC process by the #1234 standard test. The sentence, "This paper discusses the application of standard test #1234 to the measurement of product durability.", also provides data to the reader. This statement measures the attributes of the article in terms of its discussion of certain topics. If a user needs to know the percent of greater durability from process ABC, then the first sentence would be of value to him and the second sentence would be of lesser value. If, on the other hand, the user wanted to locate an article that discussed the application of standard test #1234, then the second sentence would probably be more valuable to him. Both sentences present data to the reader, independent of any value derived through use of this data.

This view of data has been formalized by Landry and Rush (11, 12). They present the following definition for the unit of data, the data element

> Define a data element, d, as the smallest thing that can be recognized as a discrete element of that class of things named by a specified attribute for a given unit of measure with a given precision of measurement.

They also define a document in terms of this definition.

> Define a document, D, as a well-ordered set of data elements.

For instance, one type of data element that the abstracting system must

identify is the sentence. Sentences provide a partitioning of the document and hence are the class of things being measured. The unit of measure is a string of characters bounded by periods. The precision is the cardinality of the set of sentences. At the same time, ADAM views each sentence as being composed of a well-ordered set of data elements, words. The unit of measure is the string of characters bounded by blanks. The precision is the cardinality of the set of words comprising a sentence.

Both a document and an abstract, which is a document itself, can be analyzed in terms of the data elements they contain. The importance of this notion lies in the fact that, since the data may be useful to a decision maker in some decision making situation, the abstract, or the parent document can be assigned a value proportional to the amount of data it contains.

6 The Relationship Between Data and Information

The value of data arises from the ability of the decision maker to use the data in making decisions. If a person were trying to decide whether to read a given document, he might first read the abstract to determine if the document would be of interest to him. If, after reading the abstract, he is able to decide either yes, he should read the document, or no, he should not; then the data has been valuable to him. We must observe his actions after reading the abstract to determine if indeed the abstract influenced his course of action. The relationship implied here between data and the decision can be seen in

the model, proposed by Yovits and Ernst (13, 14), which is shown in Figure 4.1. This model introduces the following definition:

Information is data of value in decision making.

The importance of this definition for the evaluation of abstracts is seen through the following arguments. A single data element will either be valuable or not to a decision maker in a given situation. By contrast, a set of data elements, such as an abstract, will have a value in the interval 0-1 which is proportional to the number of data elements in the set that are actually of value to the decision-maker, relative to the total number of data elements in the set. The cardinality of a set of data elements is fixed, but the cardinality of the subset that is of value to a decision maker in a particular decision-making situation is variable.

Thus, the ideal abstract should present to a user the data that will be of value to him in making his particular decision. While this is a sufficiently difficult task, this problem becomes further complicated because a single user might employ the abstract in several decision situations or many users might use the abstract in a particular decision situation. In general, an abstract will be used by many decision makers in many decision situations, so that an abstract should provide useful data to a range of decision makers over a range of decision-making situations. But it is this very variability in the use of the abstract which defeats attempts to evaluate abstracts on the basis of their usefulness to decision makers.

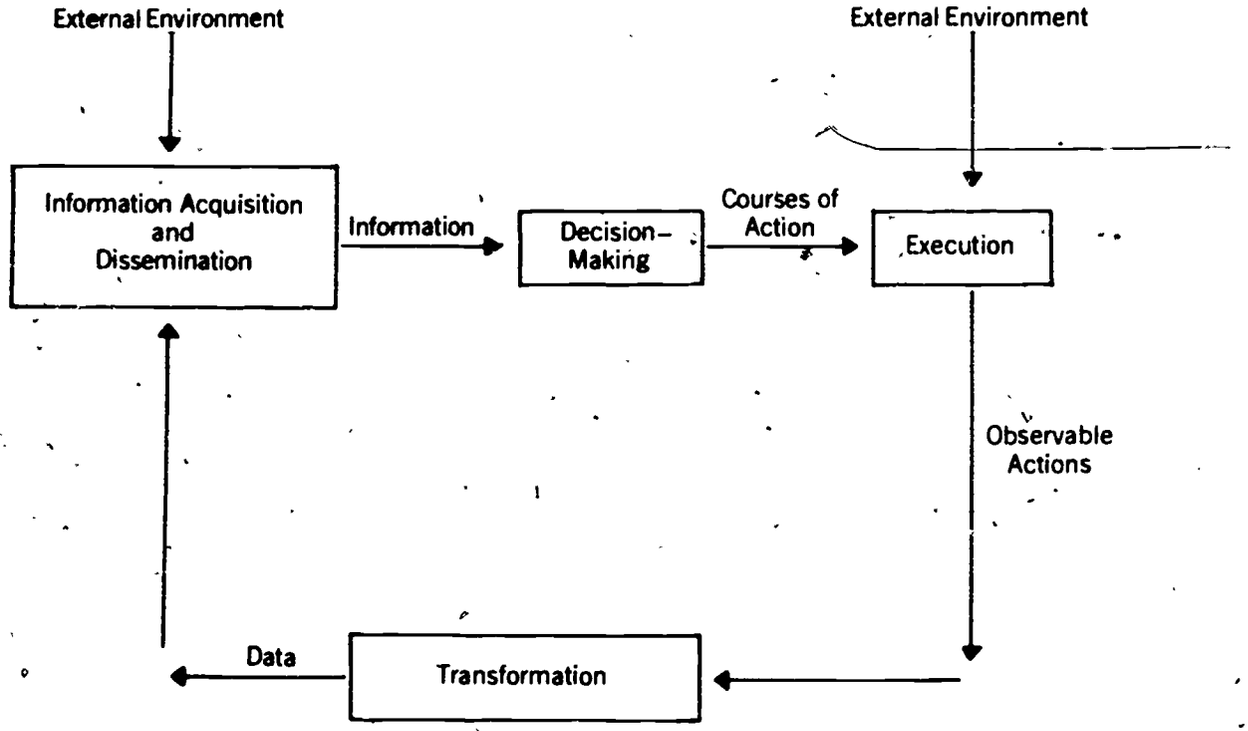


Figure 4.1 The generalized model of information flow proposed by Yovits and Ernst (13, 14).

6.1. On the Difficulty of Using "Information Content" as a Basis for Evaluation of Abstracts

Using the Yovits-Ernst model of a generalized information system, let us consider the value of a particular document in some decision-making situation. In order to measure the value of the document (set of data elements), the following essential factors must be considered:

1. The set of decision makers and their prior knowledge
2. The decision situation
3. The external environment

   (Factors 2 and 3 are assumed to be constant; this is an assumption which is impossible to verify.)
4. The execution function, that is, the decision function and the uncertainty involved with the probability of predicated outcomes
5. The isolation of just those observables that are influenced by the data contained in the document being studied.

While it would be useful to measure the information derived from an abstract and from its parent document in order to evaluate them, the list of factors which must be measured to determine the derived information makes the task of evaluation quite impractical, if not impossible.

6.2. On the Usefulness of Data Content as a Basis for Evaluation of Abstracts

It can be seen that information derivabled from a document is a relative quantity depending on the determination of value, whereas the

quantity of data contained in the document is independent of such a determination. The relationship between information and data is shown in the Venn diagram in Figure 4.2. In the Figure, the set D represents the set of data, i.e. a document. The subset of data which is valuable to a given user A is denoted I_A, the subset value to user B is denoted I_B and the subset valuable to user C is denoted I_C. The information which can be derived from a given set of data ranges from a minimum of 0, where none of the data is of value, to a maximum where all of the data is of value. Thus, the amount of information derived from a set of data can never exceed the amount of data itself.

Potential utility of the data contained in the abstract can be generally predicted based on the number of data elements contained in the abstract relative to the parent document. Since the abstract is also an abbreviated version of the parent document, it is desirable to maximize the number of data elements contained in an abstract of a particular size. The abstract which will serve the largest group of users should provide an efficient representation of the data contained in the parent document. The best abstract may not serve each of the users equally well, but we can predict that it will have the greatest likelihood of providing useful data to the largest group of users. The arguments presented in the preceding sections form the basis of an evaluation method which is described and illustrated in the remainder of this chapter.

My assumption, and one which I believe is held implicitly by most abstracting services, is that an abstract should serve a large group of

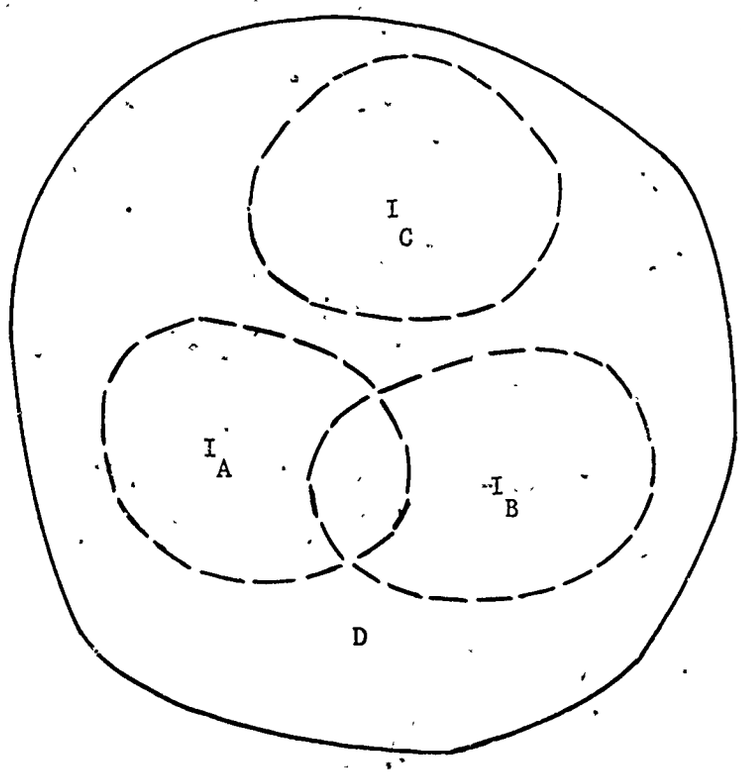


Figure 4.2 The relationship between data and information

users in a wide variety of decision-making situations. Consequently, the abstract should contain an efficient representation of the data of the parent document. While I hold that this is the best abstract for general applications, it must be realized that such an abstract may not be "best" for each individual user. Nevertheless, I believe that a good general abstract will have the greatest likelihood of providing useful data to the largest group of users in the widest variety of decision-making situations over the longest span of time.

An evaluation method, based upon the arguments presented in this and preceding sections, is described and illustrated in the remainder of this chapter.

7. A Two-Step Procedure for the Evaluation of Abstracts

In the preceding sections, I have considered the theoretical foundations for an evaluation criterion. I have argued that an abstract should present an accurate representation of the data contained in the parent document in a condensed form. I have developed a two-step objective procedure to determine how well an abstract meets this goal. The two steps of this procedure are described below.

Step 1. Determine if the abstract conforms with the criteria for an acceptable abstract.

Step 2. Determine the data coefficient for the abstract or abstracts that satisfy Step 1.

A. If there is only one abstract, then the value of the data coefficient should be greater than or equal to one.

B. If there is more than one abstract, of which at least, one has a data coefficient greater than or equal to one, the best abstract will be the one that has the highest data coefficient.

In the following sections, I will consider each of these steps in more detail.

7.1. Step 1

The first step of the evaluation procedure consists of the following test.

Determine if the abstract conforms with the criteria for an acceptable abstract.

The implementation of the first step depends on specification of the criteria for an acceptable abstract produced by a given abstracting system. Every abstract user has his own requirements for what an abstract should contain and these requirements probably vary with time. The overlap of these requirements over a large group of users is uncertain, although probably small. The Subcommittee 6 of the American National Standards Institute (ANSI) Committee Z39 has attempted to define those qualities of abstracts which should serve as uniform standards for abstract production (7). These standards are designed to be applied to the production of abstracts of documents from a wide range of subject areas. Some abstracting services have individual specifications that either coincide with or supplant the proposed ANSI standard (15).

For automatic abstracting development and evaluation research in this Laboratory, we have chosen to use the ANSI standard as described

by Weil as our goal. This standard instructs the human abstractor or computer-based abstracting system to (7)

> Keep abstracts of most papers to fewer than 250 words, abstracts of reports and theses to fewer than 500 words (preferably on one page), and abstracts of short communications to fewer than 100 words. Write most abstracts in a single paragraph. Normally employ complete, connected sentences; active verbs; and the third person. Employ standard nomenclature, or define unfamiliar terms, abbreviations, and symbols the first time they occur in the abstract.

Any abstract which does not meet this standard should be edited or rewritten to conform with these criteria.

In order to implement Step 1, there are two essential questions to be considered. First, does the abstracting system meet the design specifications and second does user feedback indicate that the design specifications should be changed. In terms of the discussion of Chapter 2, Step 1 amounts to a determination 1) of how closely the a priori and operationally defined intensions of the abstracting system correspond; 2) of how well system intension matches the user's view of their purpose.

Actual tests of which Step 1 is comprised are based upon system design specifications. This is an important point. The system has been designed to produce abstracts of a certain nature. Step 1 amounts in part to an alternative system design which should rate abstracts, produced by the original design, according to the design criteria embodied in both designs. Whenever discrepancies are found, it is necessary to determine which of the two designs gave rise to the discrepancy and to decide whether the difference is of consequence.

If so, either the abstracting system or the evaluation system must be altered to eliminate the discrepancy.

Step 1 also takes cognizance of user acceptance of the abstracts produced by the abstracting system under study and attempts to determine how to alter the system to meet the user's demands.

I shall not attempt to deal with Step 1 in detail in this dissertation. However, some observations of a general nature are given to conclude this section.

Step 1 might include a determination of whether the following criteria are satisfied by a given abstracting system:

1. Maximum length
2. Minimum length
3. Bibliographic citation format
4. Subject orientation
5. Error level
6. Style
7. Sentence completeness
8. Form (block vs. paragraphed)
9. Type (indicative, informative, etc.)
10. Timeliness

For ADAM, I will assume that the system has been designed so that the abstracts produced conform to the desired criteria. If this assumption is not valid, the system design should be modified.

There may be several abstracts which satisfy the criteria in Step 1 but which differ from one another in some respects. These differences

may be in content, organization, or sentence structure. The differences may also reflect a particular user's needs for certain topics to be included in the abstract. Thus if several abstracts of a given document pass Step 1, there is then a need to choose the best abstract from among this set. "Best" is used here in the sense of Section 6.2. The choice of best abstract is accomplished through the implementation of Step 2.

7.2. Step 2

The second step of the evaluation procedure consists of the following tests.

Determine the data coefficient for the abstract or abstracts that satisfy Step 1.

A. If there is only one abstract, then the value of the data coefficient should be greater than or equal to one.

B. If there is more than one abstract, of which at least one has a data coefficient greater than or equal to one, the best abstract will be the one that has the highest data coefficient.

The implementation of Step 2 depends upon the evaluation of a defined ratio, which I have called the data coefficient.

7.2.1. Definition of the Data Coefficient

The data coefficient is a function which expresses a relationship between the data contained in an abstract and the length of the abstract.

Since abstracts are, as I have said, abbreviated representations of the original document, they should contain as much of the data in the original document as possible while being much shorter than the original. Thus, in evaluating abstracts the data coefficient should reflect this desired property of abstracts. The following definition of the data coefficient, DC, serves this purpose.

$$DC = \frac{C}{L}$$

where C is the data retention factor,

$$C = \frac{\text{the amount of data in the abstract}}{\text{the amount of data in the document}}$$

and where L is the length retention factor,

$$L = \frac{\text{the length of the abstract}}{\text{the length of the document}}$$

Not only should the DC relate the content and length of an abstract in a meaningful way, it should make possible both comparative and absolute evaluations of abstracts. For this latter purpose, an understanding of the significance of DC values is required.

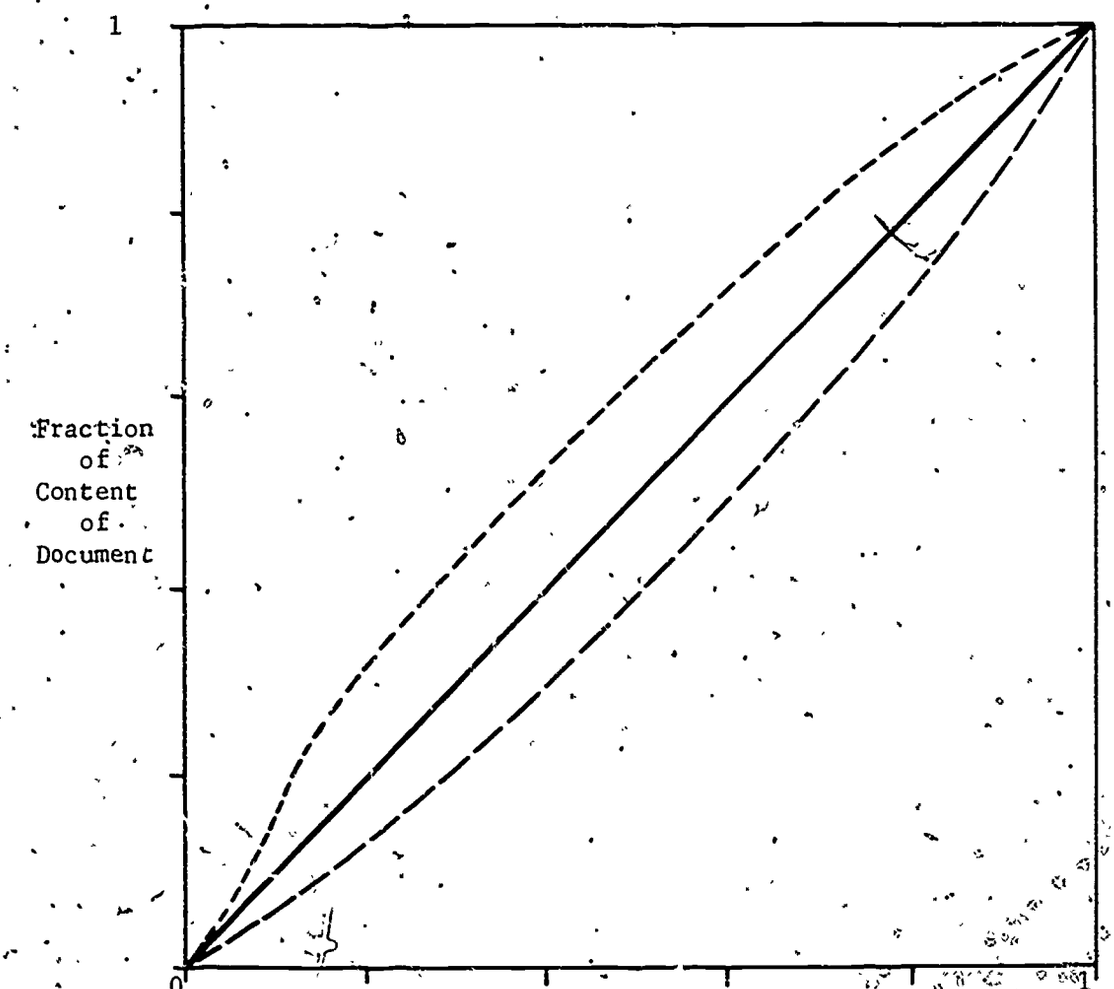
There are three classes of abstracting function:

1. The class of abstracting functions that reduce length at a greater rate than they reduce data content;
2. The class of abstracting functions that reduce length and data content at the same rate;

3. The class of abstracting functions that reduce data content at a greater rate than they reduce length.

Hypothetical curves illustrating the behavior of these three classes of abstracting functions, and representing all possible abstracts of a given document, as viewed from the perspective of the data coefficient, are given in Figure 4.3. Although smooth curves are shown, it should be emphasized that the data coefficient is not regarded as a continuous function. The reason for this, if not already obvious, will be made clear shortly. The purpose of the curves of Figure 4.3 is to establish bounds for DC values which represent acceptable abstracts. In general, we can say that abstracts whose DC values lie above the straight line will be acceptable and otherwise not. However, it is also desirable to know what significance attaches to the magnitude of the DC value. Since the magnitude of the DC value can be interpreted as the slope of a line from the origin to the point representing the DC value, we are lead to a consideration of the general shape of the curves shown in Figure 4.3.

The process of abstracting implies that there will be a reduction in length. The reduction in length can be measured in terms of the number of characters, words, sentences, or other defined units, that are retained for the abstract relative to the number of such units in the original document. It would be possible to generate a set of abstracts to represent each possible length by reducing the length of the document in unit intervals. For example, the set of abstracts could be generated by reducing the length of the document one word at



——— Represents reduction in content equal to reduction in length.

– – – – Represents reduction in content less than reduction in length.

—— —— Represents reduction in content greater than reduction in length.

Figure 4.3 Hypothetical curves for content/length comparisons in the evaluation of abstracts.

a time. By selecting words at random, it is possible to produce a set of abstracts which would provide a set of discrete points on the graph at each unit of length. The crucial question is the relationship between the reduction in length and the reduction in data content.

The relationship between data content and length for a set of possible abstracts for one document could be characterized by connecting the set of discrete points to form a curve. This curve could be analyzed to select the one best abstract, in terms of its efficient data representation, from the set. This curve could also be used to predict the behavior of the same abstracting system on another similar document. A set of such curves for a group of representative documents could be used to characterize the operation of the abstracting system being evaluated.

The construction of a set of curves would require the generation of a large number of possible abstracts. Although the task of creating the abstracts would be difficult, it is possible to experimentally determine the shape of the curves. I have performed some limited experiments by generating a set of possible abstracts for one document. The abstracts were created by selecting words at random to delete. For each word that was deleted a new abstract, shorter by one word, was formed. The data content of each of the abstracts was evaluated. Based on these experiments I have made the following observations about the nature of the curves in general. The theoretical curves can be characterized by the relationship between content and length. There are three possible values for this relationship: the fraction of the

content may be greater than the fraction of the length, C may be less than L or C may equal L. Let us consider each of these three cases.

If C > L then the abstract production method provides a greater reduction in L than C. This is clearly a desirable property because the abstract is presenting an efficient representation of the data to the user. The slope of this curve will probably be quite small for small reductions of length. Since written communication is generally quite redundant (16), it should be possible to reduce the length by eliminating some of the restatements of data elements and function words and phrases without a corresponding reduction in data. At some point though, the reduction in data may exceed the reduction in length. This will occur when the data elements are represented so efficiently that the removal of even one unit of length removes one unit of data. For example, if one unit of data is the clause which contains several words and one unit of length is the word, then the removal of the significant word of the clause, the verb, would reduce the data to zero but the length would only be reduced by one word. If the evaluation of an abstracting system showed that all abstracts produced by the system showed a greater reduction in data than in length, this would be undesirable and major system redesign would be indicated.

The critical point occurs when the percentage reduction in content is equal to the percentage reduction in length. The goal of abstracting is to reduce the length at a faster rate than the content. If the original document is so well written that it would be impossible to reduce the length without a corresponding reduction in content, then

the document would serve as its own best abstract. If the document is evaluated as an abstract of itself, the data coefficient will equal one since C and L will both be one. Therefore, there will always be at least one abstract with a data coefficient greater than or equal to one if we allow the possibility of using the document as an abstract.

Constraints may be placed on the abstracting system to produce abstracts with a specified maximum length. If, for example, the abstract must be shorter than 250 words, there is no guarantee that it will be possible to find an abstract with a data coefficient greater than one. Under constraints of this nature, the entire set of possible abstracts may have data coefficients less than one.

Based on the value of the data coefficient we can make the following generalized assessments of value:

1. If DC < 1 then the abstract is unacceptable
2. If DC = 1 then the abstract is at the minimum level of acceptability
3. If DC > 1 then the abstract is acceptable
4. If two abstracts both have DC values greater than one, then both are acceptable but the higher DC value indicates the better abstract.

If the length of the abstract is 0 then the DC is undefined. This fact does not limit the effectiveness of the DC since it is logically impossible to evaluate an abstract that does not exist. The values of the DC range from zero to the number of data elements.

The determination of the data coefficient represents the second step of the evaluation procedure mentioned earlier. It represents first, an absolute measure of abstract acceptability and second, a criterion to determine which of two abstracts is better. Since the data content provides an upper bound on the amount of information that can be derived from the data in an abstract, the data coefficient also is a prediction of the usefulness of the abstract. In order to apply the concepts for abstract evaluation to an automatic abstracting system, there must be a definition of a means a) of identifying data elements and b) of measuring length.

7.2.1.1. Definition of the Units of Data and Length

The use of data elements as a means of evaluating the quality of abstracts depends upon the identification of data elements within the abstract and within the document. The method of identification must define the boundaries of data elements in a meaningful and consistent manner. The definition of data element given earlier--a data element, d, is the smallest thing which can be recognized as a discrete element of that class of things named by a specific attribute, for a given unit of measure with a given precision of measurement--is really quite flexible. The possibility of a data element being any desired entity is a convenience from both a descriptive and a theoretical viewpoint. In order to specify a data element for a given situation, it is necessary to specify the attribute, the unit of measure, and the precision of measurement. Once the nature of the data element is specified, we can apply the data coefficient measure required by

Step 2 of the evaluation procedure I have proposed.

For application to abstract evaluation, the attribute being measured is data content. The data elements are, in this case, usually defined to be words, clauses, and/or sentences, but specific numerical quantities, equations, tables or figures could also be specified as data elements. The unit of measure could then be, for example, the word, clause, sentence, etc., and the precision of measurement would be specified by the number of data elements identified by the unit of measure. The data elements of a given abstract or document may be defined by specifying the essential properties necessary for their identification or by enumeration of all possible data elements. The manner in which a data element is specified is dependent on the function of the abstract. The data element must be defined in the context of its use as the minimum unit of data that has potential value to a decision maker.

We can measure the accuracy of representation in terms of data content. We now need a measure of the abbreviation. Perhaps the simplest measure of length is the number of characters contained in the document or abstract. The number of bits that a document occupies when in a computer could be calculated by multiplying the number of characters times the number of bits per character. (Blanks, and special symbols are included in the list of characters). Other measures of length might include the number of words, the number of sentences, or the number of printed lines. Any uniform criterion for the measurement of length would be acceptable.

7.2.1.2. Implementation of the Data Coefficient

In order to implement the data coefficient within the constraints specified above, a general formulation of the data element is required. In some of the specific applications of abstracts, as for example when abstracts are used as a source for index terms from a controlled vocabulary, it would be possible to enumerate all possible data elements, where the data element is defined as a single index term. For a general application it would be difficult to enumerate all data elements. Furthermore, the chosen data elements should be able to reflect all of the data contained in the abstract and should not be limited to a fixed list of possibilities.

A data element should, in a general sense, be defined to be the minimum unit of data that has potential value to a decision maker. In an abstract this unit of data might be expressed by a single word, but in general, several words would be needed to express one complete data element. One sentence might express one data element and a compound sentence might express two or more. Thus, it would be inadequate to merely count the number of words or the number of sentences to determine the number of data elements in an abstract.

A document conveys data to a reader 1) through the unique organization of language elements within the document and 2) through the significance imputed to these language elements outside the framework of the document. Since the significance imputed to the language elements depends on the decision maker, his prior knowledge, and his decision-making situation, we cannot ascertain the value of the data contained

in the document to all potential decision makers. But it is possible to examine the unique organization or language elements within the document and the manner in which they convey data to the user. The examination of the language elements within a document depends upon a linguistic analysis of the text to determine the structural representation of a single data element. The quality of the abstract will then be dependent on the accurate preservation of the data elements of the document.

The actual implementation of the data content criterion depends on the specification of meaningful boundaries for a data element. The identification of content elements in natural language documents has received a great deal of attention. As Christine Montgomery states (17) "... much of the recent work in linguistics, as well as in computational linguistics, might be entitled 'In search of a formalism for content representation'." The identification of concepts by means of linguistic analysis has received attention because it is significant to almost all natural language processing endeavors. None of the various linguistic research efforts have arrived at a definitive method of content representation but a few basic tenants seem to be emerging. According to Montgomery (17):

> Perhaps the most encouraging note is the emergence of certain fundamental principles to which a majority of these researchers are committed. Central among these is the notion of the predicate as pivotal in semantic and syntactic analysis. ... the term 'predicate' in this context designates any relation holding between two or more entities (its arguments in the logical sense) or any property of an entity.

Language may be considered as a means of communicating about certain

entities and the relationships between these entities.

Evidence is being found to support the preeminence of the predicate in all communication. Researchers have identified this preeminence in several different languages. This feature of linguistic analysis also appears to be independent of the subject matter of the particular example. Thus a linguistic analysis based on the centrality of the predicate could be applied to abstracts of any subject discipline or any language.

I have developed a method for identifying specific data elements. This method is based on the concept of the centrality of the predicate and on some specific ideas of Young (18). This method is designed to provide a specific means of identifying data elements where one data element is defined to be equivalent to one concept. This formulation (which is presented in the next section) is only one of several possible ways of representing data elements. Other means of defining the data element could certainly be used in the data concentration criterion. The crucial factor is that the identification and the preservation in the abstract of the data elements of the document is essential to good abstracting.

7.2.1.3. The Representation of Data Elements by Name-Relation-Name Patterns

Words and groups of words, which are called language strings, can be classified into one of two groups, names or relations. In general names are used to identify objects or constructs and relations are used to express the relationships between names. A name, denoted N, is defined as a language string assigned to a behavior imputed to an object or construct, as well as those language strings which have only

a linguistic function. A name may be simple, composite or complex. A simple name is a single word which is assigned to an object or construct. A composite name is a sequence of simple names. A complex name is an ordered triple, N_i R N_j where N_i and N_j are simple, composite, or complex names and R is a relation. Complex names allow for N - R - N patterns to appear as a name in another N - R - N pattern.

A relation, denoted R, is defined as any language string that expresses the relationship between two names. Relations may be either primary or secondary. Primary and secondary relations may be either simple or composite. Primary relations are defined as any word which functions as a verb. A secondary relation is defined as any preposition. A simple relation is a primary or secondary relation that consists of only a single word. A composite relation is a relation that is made up of a sequence of simple relations. If the sequence contains a primary relation, then it is primary; otherwise, it is secondary.

Using these definitions, it is possible to classify all words in written text as either names or relations. A sentence or part of a sentence can be described as an ordered triple, N_1 R_p N_2, where N_1 and N_2 are names (simple, composite, complex, or vacuous) and where R is a primary relation (simple or composite, but never vacuous). By categorizing words of the text into N - R_p - N patterns, it will be possible to identify data elements.

The name-relation-name pattern expresses the relationship between two objects or constructs. This formulation corresponds to the pattern of a simple sentence where the first name corresponds to the subject,

the relation to the verb, and the second name to the object. A simple sentence traditionally is used to express one idea. This formulation also presents a method for structuring information. The names may be used to represent nodes of a graph and the relations edges between the nodes. The sum of many data elements would form a complete network of data.

It is possible to identify $N - R_p - N$ data elements in a text by the following procedure. First, identify all primary relations. Second, associate with each primary relation the first name, N_1, and the second name, N_2. The names may be simple, compound, complex or vacuous. Each $N_1 - R_p - N_2$ triple, where R_p is a primary relation, is a data element. The number of data elements, which are expressed by $N_1 - R_p - N_2$ patterns will always be greater than or equal to the number of sentences in the text being examined. The number of data elements associated with various sentence constructions is given in Table 4.1. The number of data elements contained in each sentence corresponds to all basic $N_1 - R_p - N_2$ patterns.

In this definition of data element, each $N_1 - R_p - N_2$ pattern is considered to convey data to the reader in an amount equal to all other $N_1 - R_p - N_2$ patterns. Based on this assumption, all data elements in the text are given equal weights. If there were some function which could predict the potential value of certain data elements, it would be useful to rank the data elements based on this functional weighting. For example, the data elements might be given a higher weight if keywords from the title or index were present in either of the names or the

Table 4.1 Summary of the number of data elements associated with sentence constructions

	Number of Data Elements
$N - R_p - N$ Patterns	
Within sentence patterns	
Simple Sentences	
$N_1\ R_p\ N_2$	1
Compound Sentences	
$N_1\ R_{p_1}\ N_2$ 'and' $N_3\ R_{p_2}\ N_4$	2
$N_1\ R_{p_1}$ 'and' $R_{p_2}\ N_2$	2
N_1 'and' $N_2\ R_p\ N_3$	2
$N_1\ R_p\ N_2$ 'and' N_3	2
N_1 'and' $N_2\ R_p\ N_3$ 'and' N_4	4
N_1 'and' $N_2\ R_p\ N_3$ 'and' N_4, respectively	2
Complex Sentences	
Each clause, $N_1\ R_p\ N_2$	1
'that' $R_p\ N_2$	0
$N_1\ R_p$ 'that'	0
'which' $R_p\ N_2$	0
$N_1\ R_p$ 'which'	0

relation. The data elements might also be weighted based on the number of times the data element appeared in the original document. Any meaningful predictor of data element value for a given application would probably improve the performance of the evaluation measure.

8. Examples of the Application of the Evaluation Criterion

8.1. Sample Document with Six Possible Abstracts

This section provides an analysis of a sample document, its title, and its six possible abstracts in terms of the data content criterion. I will assume that each of the six possible abstracts have satisfied the criteria in Step 1. This examination will rank the six according to the criterion given in Step 2. The document and abstracts 1-5 used as material for this examination were selected because they were used as the basis for another study by Hirayama on abstract evaluation (19). The original document was written in Japanese and translated into English and is shown in Figure 4.4. Five abstracts of this document were found by Hirayama, written in Japanese, English and German. These five abstracts, prepared by different abstractors, were translated or rewritten into English by the same person. The five abstracts are shown in Figures 4.5 through 4.9. I have included a sixth abstract, produced by ADAM, and the title for examination. These are presented in Figures 4.10 and 4.11, respectively.

For the document, title, and abstract, I have identified the data elements contained in each. I have indicated the principal word in each name by a single underline and the words in each primary relation by a

Absorption Spectra and Chemical Structure. II. Solvent Effect

It was pointed out in part I that the λ_{max} of polyenes containing conjugated homochromophoric groups and auxochromic substituents may be expressed by

$$(\lambda_{max})^2 = A - BC^N \quad (1)$$

$$= a + B(1 - C^N) \quad (2)$$

where N is the homoconjugation index, which is determined by substituent and structural effects as well as by the number of conjugated chromophoric groups.

Thus, the following equation was given for alcholic solutions of polyene derivatives with auxochromic substituents

$$(\lambda^{alc}_{max})^2 = (36.98 - 39.10 \times 0.920^N) \times 10^3\ m\mu^2 \quad (3)$$

$$= \{-2.12 + 39.10 \times (1 - 0.920^N)\} \times 10^3\ m\mu^2 \quad (4)$$

We shall discuss the influence of the solvent on λ_{max}, *i.e.*, which parameters are effected by a change in solvent.

The parameters, (a, B and C) for carbon disulfide solutions of carotenoids were computed from the observed wave lengths; a and C are the same for alcohol and carbon disulfide solutions (-2.12 and 0.920, respectively), and only B is characteristic of the solvent. Formulas for the calculation of λ_{max} in various solvents were determined from the observed values with $a = -2.12$ and $c = 0.920$.

Solvent	$(\lambda max)^2$, $m\mu^2 \times 10^{-3}$	
Methanol	$(36.30 - 38.42 \times 0.920^N)$	(5)
Hexane	$(36.62 - 38.72 \times 0.920^N)$	(6)
Petroleum ether	$(36.70 - 38.82 \times 0.920^N)$	(7)
Light ligroin	$(37.05 - 39.17 \times 0.920^N)$	(8)
Chloroform	$(38.60 - 40.72 \times 0.920^N)$	(9)
Benzene	$(39.20 - 41.32 \times 0.920^N)$	(10)
Carbon disulfide	$(42.06 - 44.18 \times 0.920^N)$	(11)

The calculated values of λ_{max} in Table I are based on these formulas, which differ only with respect to B which is characteristic for each solvent; the agreement between the calculated and observed values is good except for compounds 14 in carbon disulfide, 25 in chloroform and in benzene and 27 in hexane. In the case of rhodoviolascin 26, it was assumed that $N_{methoxy} = 0$; the agreement between the observed values and those calculated with $N_{methoxy} = 0$ indicates that this substituent exerts practically no bathochromic effect.

Figure 1 shows that there is an approximately linear relationship between B^{solv} and the refractive index of the solvent.[2] The values for B^{solv} in formulas 5–11 agree well with the values calculated by the equation (see Table II)

$$B^{solv}_{calcd} = 12.976 + 19.019 \times n^{20}_D \quad (12)$$

Because the number of observations made in ether, cyclohexane and pyridine was insufficient for the determination of B^{solv}, B^{solv}_{calcd} computed from equation 12 was used in the calculation of λ_{max} in these solvents (Table III); the good agreement between the calculated and experimental values indicates the reliability of relation 12.

Total Words 321
Data Elements 18

Content 1.00
Length 1.00
DC 1.0

Figure 4.4 Sample document with data elements identified

Abstract 1.—Absorption max., of polyenes possessing auxochrome can be expressed by $(\lambda_{max.})^2 = A - BC^N$ and $\lambda_{max.}$ in various solvents can be calcd. from this formula by varying B. Approx. linear relation was found to exist between B and n of the solvent.

Total Words 39
Data Elements 3
Content .17
Length .12
DC- 1.37

Figure 4.5. Abstract 1 with data elements identified

Abstract 2.—Absorption max. of 57 kinds of polyene compds., including $H—(CH=CH)_3—H$, were measured in soln. of MeOH, hexane, petr. ether, ligroine, EtOH, $CHCl_3$, C_6H_6, and CS_2. A formula common to all these solvents can be obtained by changing B in the formula $(\lambda_{max})^2 = A + BC^n$ for relationship between $\lambda_{max.}$ and structure. There is a linear relationship between the value of B and n of various solvents.

Total Words 63
Data Elements 3
Content .17
Length .196
DC .88

Figure 4.6 Abstract 2 with data elements identified

Abstract 3.—The difference of $(\lambda_{max.})^2$, which increase as N increases, between a certain solvent and EtOH is shown to be the solvent effect of B and B has a linear relation to n_D^{20} of the solvent. Calcd. values of $\lambda_{max.}$ for 57 compds. in MeOH, hexane, Et_2O, EtOH, cyclohexane, $CHCl_3$, C_6H_6, pyridine, CS_2, petr. ether, or gasoline showed good agreement with observed values when $B_{calcd.}^{solv.} = 12.986 + 19.019\, n_D^{20}$.

Total Words 64
Data Elements 3
Content .17
Length .199
DC .84

Figure 4.7 Abstract 3 with data elements identified

Abstract 4.—The formula for calcg. the 1st absorption max. of polyenes can be expressed by $(\lambda_{max.})^2 = A - BC^N$. It is shown that the solvent modifies B and that there is a linear relationship between B and n_D^{20} of the solvent. The calcd. and observed 1st absorption max. are tabulated for 57 polyenes. The n_D^{20} and $B^{solv.}$ values were: MeOH, 1.3288 and 38.42; hexane, 1.3751 and 38.74; petr. ether, 1.37 and 38.82; Et_2O, 1.3556 and –; EtOH, 1.3633 and 39.10; light ligroine, 1.38 and 39.17; cyclohexane, 1.4264 and –; $CHCl_3$, 1.4486 and 40.72; C_6H_6, 1.5017 and 41.32; pyridine, 1.5085 and –; CS_2, 1.6319 and 44.18.

Total Words 101
Data Elements 5
Content .28
Length .31
DC .88

Figure 4.8 Abstract 4 with data elements identified

Abstract 5.—The formula given by the author for $\lambda_{max.}$ of polyenes in the form $(\lambda_{max.})^2 = a + B(1 - C^N)$ is examd. for its susceptibility to the solvent. It is shown that the parameter B is characteristic to each solvent and the following linear relationship to n_D^{20} was found: $B_{calcd.}^{solv.} = 12.976 + 19.019\ n_D^{20}$. The agreement is very good between the observed value and B value calcd. from the above formula for MeOH, hexane, petr. ether, Et_2O, ligroine, cyclohexane, $CHCl_3$, C_6H_6, pyridine, and CS_2. The parameters a and C are independent of solvents. The author calcd. $\lambda_{max.}$ for 57 polyene derivs. in the above solvents and compared them to the values given in the literature.

```
Total Words  105
Data Elements  7
Content  .39
Length  .33
DC  1.19
```

Figure 4.9 Abstract 5 with data elements identified

Title. Absorption Spectra and Chemical Structure. II.

Solvent Effect

Total Words 8
Data Elements 0
Content .0
Length .025
DC .0

Figure 4.11 Analysis of the title of the sample document

Abstract 6. A figure shows that there is an approximately linear relationship between B^{soly} and the refractive index of the solvent.

Total Words 19
Data Elements 1
Content .055
Length .059
DC .94

Figure 4.10 Abstract 6, produced by ADAM, with data elements identified

double underline. The total number of data elements and the total number of words is shown for each immediately below the text. The totals for each of the abstracts and the document are used to calculate the fraction of length and fraction of content for each abstract. These values are plotted on the graph shown in Figure 4.12.

The abstracts, document, and title can all be ranked according to the increasing value of the data coefficient. This ordering is shown in Table 4.2. Abstract 1, which represents .17 of the content in .12 of the length has the highest data concentration and should be selected as the best abstract of the six possible. Abstract 5 had the highest number of data elements of any abstract, but it represented one third of the length of the original, so its data coefficient was only second highest. Both abstracts 1 and 5 had a higher data concentration than the original document. Abstracts 6, 4, 2 and 3 showed a greater reduction in content than in length and are therefore unacceptable by the data content criterion.

This method of evaluation can be implemented to identify $N_{\cdot} - R_{p} - N_{\cdot}$ patterns by means of the text analysis programs used in Chapter V and a routine to calculate the values of the data coefficient. The most efficient method would be to utilize the structural analysis data that results from the analysis for the modification of sentences. The evaluation phase should be logically separate from the abstracting algorithm, but it need not be implemented separately.

This evaluation scheme would also provide data on the quality of the abstracts to be used as feedback to the system. For example, an

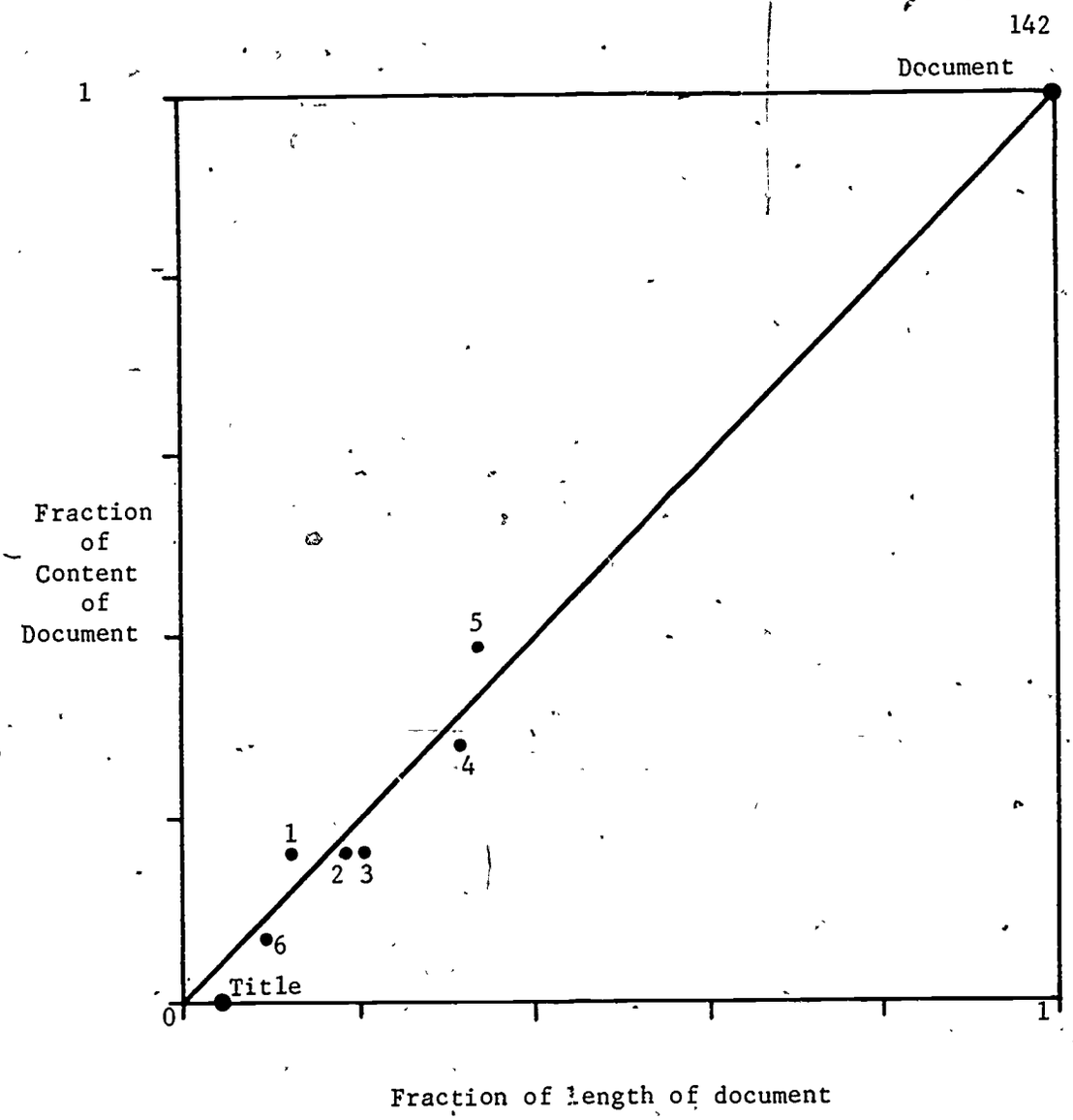


Figure 4.12 Graph of content/ length comparison for the sample document, its six abstracts, and its title.

Table 4.2 Ranking of the sample abstracts by the Data Content Criterion

Abstract	Data Coefficient
Abstract 1	1.37
Abstract 5	1.19
Document	1.00
Abstract 6	0.94
Abstract 4	0.88
Abstract 2	0.85
Abstract 3	0.84
Title	0.00

extremely long sentence that contained only one data element should be modified to make it more concise or eliminate it from the abstract. This evaluation criterion favors concise sentences with clear N - R_p - N patterns, which are the sentences that convey to the user the maximum amount of data in the minimum amount of length.

8.2. Fourteen Abstracts Produced by the Abstracting System ADAM

This section provides an evaluation, using the data concentration criterion as an absolute standard, of fourteen sample abstracts produced by ADAM. Here again I am assuming that all of the abstracts have satisfied the criteria in Step 1 of the evaluation procedure. The fourteen documents represent various subject areas, styles of writing, lengths of articles, journals, and books. The abstracts produced from these documents also vary widely in terms of quality, length, and style. The fourteen documents which I used in this study were selected from a set of documents which were keypunched by students in the introductory course in Information Storage and Retrieval at The Ohio State University. Each student in the class keypunched a document for his own investigation into the possibility of natural language processing by computer systems. The card decks which I used are duplicates of these documents which the students keypunched. The students were free to select any article they desired, so there was no control over the document selection. I choose fourteen documents that seem to present several different features and once the documents were chosen, I did not alter the sample set. (I chose to examine

fourteen documents because the weight of fourteen data decks and the output from the abstracting system was all I could carry in one trip to and from the computer center.)

The examples cited in this section serve only as an indication of the application of the data content criterion. They do not constitute a comprehensive test of the operation of the abstracting system. These examples illustrate certain trends and indicate that large-scale testing of the abstracting system (as discussed in Chapter V, Section 3.1) is warranted. I have included comments on each abstract which reflect my personal feelings on the outcome of each example.

The first example, which is shown in Figure 4.13, is an abstract of an article enticled "Water" from Chemistry. The complete document provides a discussion of some of the commonly known properties of water and of some of the recent research developments. The abstract does not indicate this aspect of the original document. The original document contains a discussion of a recent discovery, anomalous water, which is not even mentioned in the abstract. The sentences of the abstract are not related to one another and do not appear to be representative of the total content of the article. The last two sentences of the abstract do not seem to be of particular significance to the total abstract. The data coefficient of .829 reflects its low quality. The abstract omits a great deal of the significant ideas of the text and includes some extraneous verbage.

The second example is an abstract of an article from Datamation which is entitled "Automated System". The abstract, which is shown

WATER. #CHRISTOPHER HALL, CHEMISTRY 44(8), 6-10 (1971).# AFTER ALL, WATER IS THE MOST COMMON LIQUID IN THE WORLD. ONE EXPLANATION IS THAT, BECAUSE WATER IS SO WIDELY DISTRIBUTED, IF WE COULD SMELL IT WE WOULD BE SNIFFING IT ALL THE TIME, AND THAT WOULD BE A GREAT NUISANCE. TO AVOID THIS, OUR OLFACTORY MACHINERY HAS EVOLVED SO THAT IT IGNORES SUCH COMMON THINGS AS WATER. THE ANGLE H-O-H IS 104 DEGREES AND 27 MINUTES. ALSO, WATER HAS A REMARKABLE CAPACITY FOR DISSOLVING INORGANIC SALTS. THE STRENGTH OF AN H BOND IS ONLY ABOUT ONE-TENTH THE STRENGTH OF A NORMAL BOND, SUCH AS THAT BETWEEN THE H AND O ATOMS IN EACH WATER MOLECULE. EVERY O ATOM PARTICIPATES IN TWO H BONDS AND EVERY H ATOMS IN ONE. THE ARRANGEMENT IS TETRAHEDRAL-EACH O ATOM IS SURROUNDED BY FOUR OTHERS AT THE CORNERS OF A TETRAHEDRON. TWO OF THESE FOUR BONDS ARE NORMAL H-O CHEMICAL BONDS HOLDING TOGETHER INDIVIDUAL WATER MOLECULES. THE SITUATION IS REALLY LESS LIKE SHAKING THE CAREFULLY PACKED BOX OF CUBE SUGAR THAN LIKE SHAKING A DELICATELY BUILT HOUSE OF CARDS. AS THE H BONDS ARE BROKEN BY THE HEAT PUT INTO THE ICE TO MELT IT, THE WATER MOLECULES START FALLING INTO THE CAVITIES IN THE STRUCTURE. INDEED, THERE IS EVIDENCE THAT EVEN AT 100 DEGREES C, LIQUID WATER STILL CONTAINS SOME UNBROKEN H BONDS. THEORETICAL MODELS THIS PICTURE OF LIQUID WATER, THE COLLAPSED ICE-STRUCTURE MODEL, UNDOUBTEDLY HAS A LOT OF TRUTH IN IT. THIS MODEL WAS THOUGHT UP BY J. D. BERNAL AND R. H. FOWLER IN 1933 WHILE THEY WERE FOGBOUND AT A MOSCOW AIRPORT. THEY WERE LED TO THEIRCONCLUSION BY STUDYING THEIR RECENTLY OBTAINED X-RAY DIFFRACTION PATTERN OF WATER WHICH SHOWED STRONG RESEMBLANCES TO THAT OF ICE.

Document	Abstract
Total Words 2926	Total Words 287
Data Elements 283	Data Elements 23
	Content .0813
	Length .0981
	DC .829

Figure 4.13 Computer-produced abstract of "Water"

in Figure 4.14, represents .205 of the length of the original article. The article reports several developments in the field of system design. These developments are surveyed in the abstract. The abstract tends to use too many words to express the concepts presented. This abstract, although it has a data coefficient of .926, could be edited to make it more concise. If the abstracts produced by the system are consistently too long but still contain a high fraction of the data, then it might be feasible to include an editing step to follow the abstracting operation.

The third example, shown in Figure 4.15 presents an abstract of an article on "Pumped Storage" from the Ohio State Engineer. The Ohio State Engineer is a magazine which is produced at Ohio State and has a limited distribution. The articles tend to be short and not too technical. The article on "Pumped Storage" is a typical article and is aimed at an audience of undergraduate engineers. The abstract for this article provides an accurate representation of the content and style of the article. The data coefficient is 1.088 and indicates that the abstract is acceptable according to the criterion.

The fourth example, shown in Figure 4.16 is an abstract of a chapter from B. A. Trakhtenbrot's book, Algorithms and Automatic Computing Machines. The abstracting system was not designed originally to abstract chapters from books, but this abstract provides an excellent representation of the content of this section of the book. The chapter deals with the need for and the development of a precise definition of the word "algorithm". In reading the abstract, it seems to omit the

AUTOMATED SYSTEM. #ROBERT V. HEAD,DATAMATION 17(16), 22-24 (1971).# PERHAPS THE MOST NOTABLE ACHIEVEMENT TO DATE HAS BEEN THE DEVELOPMENT AND USAGE OF SYSTEM SIMULATORS LIKE GPSS, SCERT AND CSS TO ASSIST THE SYSTEM DESIGNER IN CONFIGURING TODAY'S COMPLEX SYSTEMS. WITHOUT SUCH SIMULATORS, RELIANCE ON ANALYTICAL METHODS WOULD LEAVE THE SYSTEM DESIGNER MUCH MORE VULNERABLE TO POTENTIALLY DISASTROUS THROUGHPUT ESTIMATING ERRORS. FORMATTED FILE ORGANIZATION. SEVERAL STUDIES HAVE BEEN CONDUCTED BY IBM FOR THE AIR FORCE WITH THE OBJECTIVE OF IMPROVING THE DESIGN OF FILES WHICH OPERATE UNDER THE FORMATTED FILE SYSTEM, A DATA MANAGEMENT PACKAGE WIDELY USED BY U.S. GOVERNMENT AGENCIES. THIS WORK HAS INCLUDED THE CONSTRUCTION OF TWO FILE STRUCTURE SIMULATION MODELS TO AID THE SYSTEM DESIGNER IN DEALING WITH COMPLEX DATA STRUCTURES. THE FIRST OF THESE, FOREM 1, EMBODIED ANALYTICAL TECHNIQUES WHICH, WHILE VERY FAST, EXHIBITED SEVERAL DEFICIENCIES; APPLICATION SYSTEM GENERATOR. THERE HAVE BEEN NUMEROUS EFFORTS BY COMPUTER MANUFACTURERS AND SOFTWARE COMPANIES TO PERFECT GENERALIZED APPLICATION PACKAGES FOR SUCH FUNCTIONS AS PAYROLL, ACCOUNTS RECEIVABLE, AND INVENTORY CONTROL. ESSENTIALLY, THESE PACKAGES HAVE THE OBJECTIVE OF NOT MERELY EASING THE ANALYST'S JOB BUT ACTUALLY ELIMINATING IT, AT LEAST IN COMMONLY ENCOUNTERED APPLICATION AREAS. EXPERIENCE HAS SHOWN, THOUGH, THAT WHILE PACKAGES FOR APPLICATIONS LIKE PAYROLL HAVE BEEN SUCCESSFULLY IMPLEMENTED, A GREAT DEAL OF CUSTOM TAILORING BY ANALYSTS AND PROGRAMMERS IS OFTEN REQUIRED. PACKAGES CAN HELP TO FREE THE COMPANY SYSTEMS STAFF FROM CONCERN WITH MUNDANE PROCESSING PROBLEMS, BUT THEY MAKE LITTLE CONTRIBUTION TO IMPROVED SYSTEM DESIGN METHODOLOGY. THE APPROACH TO APPLICATIONS TAKEN BY IBM WITH ITS SMALL-SCALE SYSTEM/3 COMPUTER DOES, HOWEVER IBM HAS MADE AVAILABLE FOR CERTAIN BASIC APPLICATIONS, LIKE PAYROLL, AN APPLICATION CUSTOMIZER SERVICE. BY COMPLETING A DETAILED QUESTIONNAIRE FOR EACH APPLICATION, THE USER INDICATES WHICH PROCESSING METHODS HE REQUIRES FROM AMONG THOSE AVAILABLE. THESE FORMS ARE THEN PROCESSED AT AN IBM SYSTEMS CENTER WHICH PRODUCES A SET OF PROGRAMMING AIDS FOR ALL PROGRAMS, CARD RECORDS AND DATA FIELDS REQUIRED BY THE APPLICATION. THE PROGRAMMER THEN USES THESE AS THE BASIS FOR PROGRAM CODING. THIS PROCEDURE INVOLVES A PACKAGE IN THE SENSE THAT A MODEL SYSTEM DESIGN IS MADE AVAILABLE TO THE USER. WHILE THE DESIGN ITSELF CANNOT BE CHANGED BY THE USER, NUMEROUS OPTIONS ARE AVAILABLE TO HIM, AND THE LEAD TIME FOR SYSTEM DEVELOPMENT IS REDUCED.

Document	Abstract
Total Words 1777	Total Words 364
Data Elements 116	Data Elements 22
	Content .1897
	Length .2048
	DC .926

Figure 4.14 Computer-produced abstract of "Automated System"

PUMPED STORAGE. #RICHARD FRENCH,OHIO STATE ENGINEER,53(3),JAN,1971,PG4,5,24#
TOPOGRAPHICALLY, THE ROCKY MOUTAIN STATES ARE IDEAL FOR PUMPED STORAGE;
UNDOUBTEDLY, THE SUCCESS OF THE CHICAGO DEEP PROJECT WILL AFFECT THIS IDEA. PURE
PUMPED STORAGE IS HYDROLOGICALLY INDEPENDENT; PUMPED STORAGE PLANTS ARE
MECHANICALLY SIMPLE AND VERY RUGGED; PUMPED STORAGE PLANTS ARE ALSO
SELF-MODERNIZING. WITH THE ADVENT OF MORE EFFICIENT NUCLEAR UNITS AND WITH THE
DEVELOPMENT OF GAS TURBINES, THE COST OF PUMPED STORAGE POWER WILL DROP; I.E.,
FOR EACH GAIN IN THERMAL POWER EFFICIENCY THERE IS AN EQUAL GAIN IN PUMPED
STORAGE EFFICIENCY. SINCE 1928 AND THE ROCKY RIVER PLANT, PUMPED STORAGE
CAPACITY HAS INCREASED BY MORE THAN 800%. ALONG WITH THIS GREAT UPSURGE HAS COME
COST CUTTING DEVELOPMENTS.

Document	Abstract
Total Words 1277	Total Words 110
Data Elements 96	Data Elements 9
	Content .0938
	Length .0861
	DC 1.088

Figure 4.15 Computer-produced abstract of "Pumped Storage"

THE NEED FOR A MORE PRECISE DEFINITION OF "ALGORITHM". #B. A. TRAKHTENBROT, ALGORITHMS AND AUTOMATIC COMPUTING MACHINES, 52-57(1963).# AS WE HAVE ALREADY SEEN, THE ACTUAL APPLICATION OF AN ALGORITHM MAY TURN OUT TO BE VERY LENGTHY, AND THE JOB OF RECORDING ALL OF THE INFORMATION INVOLVED MAY BE ENORMOUS. UNTIL RECENTLY, THERE WAS NO PRECISE DEFINITION OF THE CONCEPT "ALGORITHM" AND THEREFORE THE CONSTRUCTION OF SUCH A DEFINITION CAME TO BE ONE OF THE MAJOR PROLBEMS OF MODERN MATHEMAICS. IT IS VERY IMPORTANT TO POINT OUT THAT THE FORMULATION OF A DEFINITION OF "ALGORITHM" MUST BE CONSIDERED NOT MERELY AN ARBITRARY AGREEMENT AMONG MATHEMATICIANS AS TO WHAT THE MEANING OF THE WORD "ALGORITHM" SHOULD BE. THE DEFINITION HAS TO REFLECT ACCURATELY THE SUBSTANCE OF THOSE IDEAS WHICH ARE ACTUALLY HELD, HOWEVER VAGUELY, AND WHICH HAVE ALREADY BEEN ILLSUTRATED BY MANY EXAMPLES. WITH THIS AIM, A SERIES OF INVESTIGATIONS WAS UNDERTAKEN BEGINNING IN THE 1930'S FOR CHARACTERIZING ALL THE METHODS WHICH WERE ACTUALLY USED IN CONSTRUCTING ALGORITHMS. THE PROBLEM WAS TO FORMULATE A DEFINITION OF THE CONCEPT OF ALGORITHM WHICH WOULD BE COMPLETE NOT ONLY IN FORM BUT MORE IMPORTANT, IN SUBSTANCE. VARIOUS WORKERS PROCEEDED FROM DIFFERENT LOGICAL STARTING POINTS. AND BECAUSE OF THIS, SEVERAL DEFINITIONS WERE PROPOSED. IT TURNED OUT THAT ALL OF THESE WERE EQUIVALENT, AND THEY DEFINED THE SAME CONCEPT.

Document	Abstract
Total Words 2257	Total Words 203
Data Elements 157	Data Elements 16
	Content .1019
	Length .0907
	DC 1.124

Figure 4.16 Computer-produced abstract of "The Need for a More Precise Definition of 'Algorithm'"

definition of "algorithm". This is understandable because, in fact, the definition is actually given in a subsequent chapter of the book and not in the text being abstracted. The abstract has a data coefficient of 1.124 and represents .091 of the length of the original. In this specific case, ADAM is shown to have applicability to more than just journal articles.

The fifth example, shown in Figure 4.17, is another case where the abstracting program provided surprisingly good results. The abstract is of an article entitled "The Clavichord and How to Play It" which appeared in Clavier. The abstract provides an accurate representation of the content of the original with sentences selected from both the description of the instrument and the techniques for playing it. The original article used several quotations from musicians to elaborate main points of the article. All but one of these quotations were not included in the abstract. The one quotation that was included expresses "the especially remarkable features of clavichord music", an idea central to the article. This article had a data coefficient of 1.450, the second highest of the set of sample abstracts.

The sixth example, shown in Figure 4.18 provides .140 of the content in .147 of the length of the original document. The abstract provides an accurate representation of the content of the article "Solving Artwork Generation Problems by Computer" which appeared in Electro Technology. This abstract might be improved by editing it to make it more concise. Its data coefficient of .953 reflects its general mediocrity.

THE CLAVICHORD AND HOW TO PLAY IT. #MARGERY HALFORD, CLAVIER 9(2), 38-41 (1970).# ESSENTIALLY, THE CLAVICHORD IS A SHALLOW RECTANGULAR BOX WHOSE FRAGILE STRINGS, UNDER LIGHT TENSION, ARE STRUNG HORIZONTALLY FROM A SINGLE BRIDGE OVER A THIN SOUNDBOARD. THE KEYS ARE SIMPLE LEVERS WITH A BRASS BLADE CALLED A TANGENT MOUNTED VERTICALLY ON THE FAR END. THE SOUND PRODUCED IS EXTRAORDINARILY RICH IN OVERTONES. THE TONE OF THE CLAVICHORD DOES NOT EXIST READY-MADE AS IT DOES ON THE PIANO AND HARPSICHORD; IT IS FORMED AND SHAPED BY THE FINGER, AS ON A BOWED STRINGED INSTRUMENT, WITH THE RESULT BEING A GENUINE, DIRECT, LIVING "FEEL OF THE STRING" . AS LONG AS HIS FINGER REMAINS IN CONTACT WITH THE KEY, THE PLAYER RETAINS CONTROL OF THE SOUND. THE CLAVICHORD IS THE LEAST MECHANIZED AND THE MOST RESPONSIVE OF ALL KEYBOARD INSTRUMENTS IN THAT IT MEETS THE PLAYER HALFWAY IN ITS INSTANT AND FAITHFUL TRANSMISSION OF HIS SLIGHTEST MUSICAL INTENTIONS. EMBELLISHMENTS CAN BE PLAYED CRISPLY AND BRILLIANTLY. SHAKES, SNAPS, APPOGGIATURAS, TRILLS, TURNS, MORDENTS, AND SLIDES -- ALL SO CHARACTERISTIC OF THE PERIOD WHEN THE CLAVICHORD ENJOYED ITS GREATEST POPULARITY — ARE IDEALLY SUITED TO THE INSTRUMENT'S EXQUISITE CLARITY AND RICHNESS OF TONE. THE ACTION IS SHALLOW AND VIRTUALLY WEIGHTLESS. IT IS A PHENOMENON OF THE DOUBLE-ENDED LEVER THAT THE TONE PRODUCED BY A STRIKING FORCE WILL SOUND BETTER, SWEETER, AND RICHER AT MAXIMUM LEVER LENGTH. FOR THIS REASON, THE KEYS OF THE CLAVICHORD ARE PLAYED AS NEAR TO THE FRONT EDGES AS POSSIBLE. EXCEPT FOR THE PLAYING OF OCTAVES, THE THUMB IS NEVER USED ON A RAISED KEY; DISPLAY PIECES OF A VIRTUOSO CHARACTER ARE GENERALLY UNSUITED TO THE PERSONAL QUALITIES OF THE CLAVICHORD. CRAMER SAYS THAT THE ESPECIALLY REMARKABLE FEATURES OF CLAVICHORD MUSIC ARE "FLUIDITY, SUSTAINED MELODY DIFFUSED WITH EVER-VARYING LIGHT AND SHADOW, THE USE OF CERTAIN MUSICAL SHADING AND ALMOST COMPLETE ABSTINENCE FROM PASSAGES WITH ARPEGGIOS, LEAPS, AND BROKEN CHORDS;

Document	Abstract
Total Words 1728	Total Words 283
Data Elements 139	Data Elements 33
	Content .2374
	Length .1638
	DC 1.450

Figure 4.17 Computer-produced abstract of "The Clavichord and How to Play it"

SOLVING ARTWORK GENERATION PROBLEMS BY COMPUTER. #KENNETH SANDERSON, ELECTRO TECHNOLOGY 84(5), 63-65 (1969).# COMPUTER-AIDED DESIGN TECHNIQUES NOW CAN PRODUCE AUTOMATICALLY MASK-ARTWORK FROM PUNCHED-CARD INPUTS WITH UP TO 3:1 SAVINGS IN TIME AND COST. CHIP ARTWORK IS PRODUCED AT 10 X. THE 10 X CHIP ARTWORK THUS GENERATED IS DIRECTLY MOUNTED IN A STEP-AND-REPEAT CAMERA TO PRODUCE A 1X PLATE OF A WAFER LEVEL OF MANY CHIPS. THERE ARE SEVERAL METHODS OF ENTERING THE GEOMETRIC DESCRIPTION OF A MASK INTO THE COMPUTER. ONE METHOD IS TO DESCRIBE THE GEOMETRIES WITH THE AID OF A DESIGN LANGUAGE PUNCHED ON CARDS. A SECOND METHOD IS TO ENTER DATA BY MEANS OF AN INTERACTIVE GRAPHICAL DISPLAY SYSTEM. EACH METHOD HAS ADVANTAGES AS WELL AS LIMITATIONS. DATA ENTRY VIA PUNCHED CARDS ALLOWS MORE ENGINEERS TO USE THE SAME SYSTEM IN THE SAME PERIOD OF TIME. IT IS ALSO TRUE THAT THE DEFINITION OF SPECIFIC DIMENSIONS IS MORE EASILY ACCOMPLISHED WHEN NUMBERS ARE PUNCHED INTO CARDS, INSTEAD OF ADJUSTING A LIGHT PEN TRACKING SYMBOL TO A SPECIFIC COORDINATE. A FURTHER LIMITATION TO CRT SCREEN ENTRY IS THE NEED FOR SPECIAL DEVICES OR ADAPTATIONS TO ENABLE THE ENGINEER TO SPECIFY THE PRECISE DIMENSIONS THAT HE REQUIRES. TO RETAIN THE SPECIAL ADVANTAGES OF BOTH METHODS AND TO AVOID THE RESTRICTING LIMITATIONS, A COMMON DATA BASE IS USED FOR INTERNALLY DESCRIBING THE GEOMETRIES OF MASK ARTWORK. THIS BASE ADDS SUBSTANTIALLY TO THE CAPABILITY/FLEXIBILITY OF THE COMPUTER AIDED ARTWORK GENERATION SYSTEM.

Document	Abstract
Total Words 1390	Total Words 234
Data Elements 107	Data Elements 15
	Content .1402
	Length .1472
	DC 0.953

Figure 4.18 Computer-produced abstract of "Solving Artwork Generation Problems by Computer"

The seventh example is the abstract of "Magnetic Bubbles" shown in Figure 4.19. The article appeared in Scientific American and contained 5068 total words to make it the longest of the sample documents. The sentences selected for the abstract reflect the content of the entire article. The abstract, which is .187 of the length of the original, is too long to be written as a single paragraph. The abstract could be improved by either editing it or adding procedures to enable the abstracting system to paragraph the output. The data coefficient value of 1.136 reflects the fact that although the abstract is long, it contains .212 of the data as I have measured it.

The eighth example, shown in Figure 4.20, is the abstract of "Educational Decision Making" which appeared in Today's Education. The title is essential to the preservation of the intended meaning of the document in the abstract. The first several sentences in the abstract do not clearly point out that the type of decision making being discussed refers to the nation's educational system. The abstract, although not in error, is not as clearly stated as it might be. The data coefficient of .958 reflects the need to improve this abstract somewhat.

The ninth example is taken from Psychology Today and appears in Figure 4.21. The article, entitled "Families can be Unhealthy for Children and Other Living Things", presents a strong theme with supporting evidence. The theme of the article can be best summarized by the sentence which appears first in the abstract, "The myth of the family blinds us to the dangers of our normal child-care practices". This sentence indicates the strong opinions of the author. The article

MAGNETIC BUBBLES. #ANDREW H. BOBECK, H. E. D. SCOVIL, SCIENTIFIC AMERICAN 224(6),78-90(JUNE,1971).# NOW LET US LOOK AT A THIN WAFER CUT FROM A SPECIALLY SYNTHESIZED SINGLE CRYSTAL OF MAGNETICALLY ANISOTROPIC MATERIAL. WHEN THE WAFER IS VIEWED BY POLARIZED LIGHT ONE SEES A PATTERN OF WAVY STRIPS REPRESENTING DOMAINS. IN HALF OF THE STRIPS THE TINY INTERNAL MAGNETS POINT UP, IN THE OTHER HALF THEY POINT DOWN. DEPENDING ON THE ORIENTATION OF THE POLARIZING FILTER, ONE SET OF THE STRIPS WILL LOOK BRIGHT AND THE OTHER DARK. THE TWO SETS OF STRIPS OCCUPY EQUAL AREAS. NEXT LET US IMMERSE THE WAFER IN AN EXTERNAL MAGNETIC FIELD PERPENDICULAR TO THE WAFER AND OBSERVE WHAT HAPPENS WHEN WE SLOWLY INCREASE THE STRENGTH OF THE FIELD. AS THE STRENGTH IS RAISED THE WAVY STRIPS WHOSE MAGNETNETIZATION IS OPPOSED BY THE FIELD BEGIN TO GET NARROWER. THE PROCESS CONTINUES INTO THE SMALL CIRCLES WE CALL BUBBLES. THE BUBBLES ARE ACTUALLY CYLINDERS SEEN END ON. RAISING THE EXTERNAL FIELD STILL FURTHER CAUSES THE BUBBLES TO SHRINK UNTIL FINALLY THEY DISAPPEAR ALTOGETHER. DEPENDING ON THE MATERIAL, THE BUBBLES HAVE A DIAMETER RANGING FROM A FEW MICRONS TO SEVERAL HUNDRED. THEY ARE STABLE OVER A THREE-TO-ONE RANGE IN DIAMETER. EACH BUBBLE ACTS LIKE A TINY MAGNET AFLOAT IN THE SEA OF A MAGNETIC FIELD OF OPPOSITE POLARITY. THE EXTREME MOBILITY OF THE BUBBLE CAN BE DEMONSTRATED BY MOVING A FINE MAGNETIZED WIRE ACCROSS THE SURFACE OF THE WAFER WHILE OBSERVING THROUGH THE MICROSCOPE HOW THE BUBBLES CAN BE PUSHED EFFORTLESSLY IN ANY DIRECTION. AT THE SAME TIME THE BUBBLES REPEL ONE ANOTHER AND MAINTAIN A FAIRLY UNIFORM SPACING BECAUSE THEY ARE ALL SIMILARLY POLARIZED. ONE CAN SHOW THAT IF THE EXTERNAL FIELD THAT PRODUCES THE BUBBLES IS HELD CONSTANT WITHIN A RANGE OF PLUS OR MINUS TWENTY PERCENT, THE BUBBLES ARE COMPLETELY STABLE AND CAN BE MOVED ABOUT INDEFINITELY. THUS WE HAVE DUPLICATED ON A MICROSCOPIC SCALE OBJECTS AS DURABLE AND AS IMPENETRABLE AS BILLIARD BALLS, WITH THE ADDED ADVANTAGE THAT THEY REPEL ONE ANOTHER. MOREOVER, BUBBLES CAN BE CREATED ANYWHERE THEY ARE DESIRED, AND THEY CAN BE DESTROYED BY TECHNIQUES WE SHALL DESCRIBE BELOW. THE FIRST MAGNETIC MATERIALS FOUND TO HAVE THE DESIRED PROPERTIES FOR STUDYING THE NEW BUBBLE TECHNOLOGY WERE ORTHOFERRITES, A SPECIAL CLASS OF FERRITES WITH THE CHEMICAL FORMULA RFEO3, WHERE R REPRESENTS YTTRIUM OR ONE OR MORE REARE-EARTH ELEMENTS. OTHER CRYSTAL-GROWING METHODS HAVE ALSO BEEN STUDIED, AND RECENTLY GOOD WAFERS HAVE BEEN CUT FROM SINGLE-CRYSTAL RODS PULLED DIRECTLY FROM THE MELT. IN THE MOST SATISFACTORY GARNET SAMPLES THE BUBBLE DIAMETER IS ABOUT THREE MICRONS, WHICH ALLOWS THE PACKING IN OF A MILLION BUBBLES PER SQUARE INCH. BY WRAPPING THE WIRE WITH DRIVING COILS IT WOULD BE POSSIBLE TO MOVE THE MAGNETIC SLUGS THROUGH THE WIRE AT HIGH SPEEDS, MUCH AS OIL IS PUMPED THROUGH A PIPELINE. *ONE IMPORTANT DRAWBACK WAS THAT NO PRACTICAL WAY COULD BE FOUND TO MOVE SLUGS BETWEEN WIRES EXCEPT BY READING A SLUG OUT OF ONE WIRE AND WRITING IT INTO ANOTHER, THE DOMAIN WALLS AT EACH END OF A MAGNETIC SLUG DO NOT CUT THROUGH THE WIRE AT RIGHT ANGLES BUT EXTEND FORE AND AFT IN THE SHAPE OF TWO LONG CONES. CONTROL REQUIRES THE CREATION OF MAGNETIC DRIVING FIELDS, MAGNETIC FIELDS WITH COMPONENTS IN-THE-PLANE OF THE WAFER. TWO GENERAL METHODS ARE AVAILABLE. THE FIRST METHOD IS CALLED CONDUCTOR ACCESS. THE SECOND METHOD, CALLED FIELD ACCESS, INVOLVES IMMERSING THE ENTIRE WAFER IN EITHER A PULSATING OR A ROTATING MAGNETIC FIELD THAT ACTS ON THE BUBBLES BY MEANS OF CAREFULLY PLACED SPOTS OF MAGNETIC MATERIAL THAT CONCENTRATE THE FIELD. IF AN ORDINARY BINARY CODE IS USED, A BUBBLE STANDS FOR ONE AND THE ABSENCE OF A BUBBLE STANDS FOR ZERO. BUBBLES CAN READILY BE MOVED IN TWO DIMENSIONS BY ADDING A SECOND SET OF LOOPS AT RIGHT ANGLES TO THE FIRST. THE TROUBLE WITH CONDUCTOR METHODS IS THAT A GREAT MANY ACCURATELY PLACED CONDUCTORS WHOSE DIMENSIONS ARE COMPARABLE TO THE SIZE OF BUBBLES MUST BE INTERCONNECTED WITH EXTERNAL-ACCESS CIRCUITS. THIS PROBLEM IS GREATLY SIMPLIFIED BY THE FIELD-ACCESS APPROACH. ONE FIELD-ACCESS METHOD INVOLVES RHYTHMICALLY RAISING AND LOWERING THE OVERALL MAGNETIC BIAS ON THE WAFER SO THAT BUBBLES ALTERNATELY SHRINK AND EXPAND. UNDER THE INFLUENCE OF THE ROTATING MAGNETIC FIELD A NEW BUBBLE WILL BE PRODUCED FOR EVERY COMPLETE REVOLUTION OF THE FIELD, AND IT WILL BE PROPAGATED TO THE RIGHT. IT IS ALSO A SIMPLE MATTER TO DIVIDE ANY STREAM OF BUBBLES BY TWO BY CREATING A LITTLE TRAP, OR BYPASS, THAT SHUNTS EVERY OTHER BUBBLE TO ONE SIDE AND REMOVES IT FROM THE MAINSTRAM. ELECTRONIC TELEPHONE-SWITCHING SYSTEMS CLOSELY RESEMBLE LARGE DIGITAL COMPUTERS. REPLICATION, WHICH IN A CONVENTIONAL SEMICONDUCTOR LOGIC SYSTEM IS CALLED "FANOUT," IS THE DUPLICATION OF EXISTING BINARY STATES. SINCE BINARY DATA ARE GENERALLY CONSUMED WITHIN CALCULATION CENTERS, THE ABILITY TO REPLICATE THE DATA FOR FUTURE MANIPULATIONS IS ESSENTIAL. TO MAKE REPLICATION POSSIBLE IN OUR BILLIARD-BALL MODEL THE FIRST MODIFICATION REQUIRED IS TO DIVIDE THE BASIC CELL INTO TWO CELLS, ONE ABOVE THE OTHER. STATES CORRESPONDING TO ONE OR ZERO ARE SYMMETRICAL AND ARE DEFINED ARBITRALLY. WE HAVE BEEN MOST SUCCESSFUL WITH FIELD ACCESS BY ROTATING AN IN-PLANE FIELD WITH A STRUCTURE CONSISTING OF "T'S" AND BARS OR SIMILAR SHAPES. TO ENTER DATA WE SELECTIVELY TRANSFER BUBBLES FROM A BUBBLE RESERVOIR, WHICH IS IN REALITY A MINOR LOOP EQUIPPED WITH A BUBBLE-GENERATOR AT ONE END AND A BUBBLE-EATER AT THE OTHER. WE ANTICIPATE THAT MAGNETIC BUBBLES WILL PROVIDE LARGE-CAPACITY INFORMATION STORAGE OF HIGH RELIABILITY AT VERY LOW COST.

Document	Abstract	
Total Words 5068	Total Words 946	Content .2121
Data Elements 363	Data Elements 77	Length .1867
		DC 1.137

Figure 4.19 Computer-produced abstract of "Magnetic Bubbles"

EDUCATIONAL DECISION MAKING. #WILLIAM L. PHARIS, JOHN C. WALDEN, LLOYD E. ROBISON, TODAY'S EDUCATION 58(7),52-54(1969). # PEOPLE WHO TRADITIONALLY HAVE HAD LITTLE OR NO VOICE IN DECISION MAKING ARE UNEQUIVOCALLY STATING, "WE WILL BE INCLUDED." SOME CRITICS WOULD MODIFY AND OTHERS WOULD DESTROY THE PRESENT STRUCTURE AS A PRELUDE TO CREATING A NEW FRAMEWORK. ESSENTIALLY, THIS MONSTER HAS TWO HEADS: THE LEGAL SYSTEM CONSISTING OF THE FORMAL GOVERNMENTAL BODIES AND OFFICIALS AT FEDERAL, STATE, AND LOCAL LEVELS WHO EXERCISE CONSTITUTIONAL, STATUTORY, AND JUDICAL AUTHORITY IN REGARD TO EDUCATION AND THE EXTRALEGAL NETWORK COMPOSED OF THOSE PERSONS, GROUPS, AND ORGANIZATIONS THAT ARE NOT PART OF THE FORMAL, LEGAL STRUCTURE BUT THAT DO INFLUENCE ITS DECISION-MAKING PROCESS. THE TWO ARE INTERDEPENDENT AND CONSTANTLY INTERACTIVE. PROPONENTS OF SPECIFIC AID PROGRAMS MAINTAIN THAT SUCH ASSISTANCE WILL NOT DESTROY STATE AND LOCAL AUTHORITY. THEY ARGUE THAT STATE AND LOCAL OFFICIALS CAN REFUSE FEDERAL AID IF THEY SO DESIRE AND, MOREOVER, THERE ARE OPTIONS WITHIN ANY FEDERAL AID PROGRAM WHICH PERMIT LOCAL OFFICIALS TO MAINTAIN THE INTEGRITY OF THEIR OWN EDUCATIONAL PROGRAMS. THE SCHOOL BOARD MUST TAKE THE FINAL, FORMAL ACTION TO LEGALIZE A DECISION, BUT IN DETERMINING THE PROCESS BY WHICH DECISIONS WILL BE REACHED, THE BOARD MAY CONSULT WITH OTHER GROUPS OR EVEN VOLUNTARILY ENTER INTO COLLECTIVE BARGINING AGREEMENTS WITH OTHER GROUPS. IN NO WAY DOES A BOARD OF EDUCATION ABROGATE ITS STATUTORY RESPONSIBILITY BY PROVIDING A MEANS THAT ENABLE OTHER GROUPS TO HAVE A VOICE IN MAKING DECISIONS. THE SCHOOL BOARD, RECOGNIZING THAT SELECTION OF STAFF IS A PROFESSIONAL DECISION, COULD THEN RATIFY THE TEACHERS' SELECTIONS. IN SUM, THE MOST IMPORTANT FACTOR AMONG A NUMBER OF THOSE AFFECTING THE LEGAL STRUCTURE FOR DECISION MAKING WILL BE THE POLITICAL DECISIONS MADE BY THE AMERICAN BODY POLITIC DURING THE 1970'S. PROFESSIONAL EDUCATORS WILL PLAY A ROLE IN MAKING THOSE POLITICAL DECISIONS, AND IN SO DOING THEY MUST ADDRESS THENSELVES TO FUNDAMENTAL QUESTIONS REGARDING THE FORMAL DECISION-MAKING PROCESS. FOR LACK OF A BETTER TERM, THIS HUGE GROUP MIGHT BE CALLED THE "GENERAL PUBLIC" AND THEIR INTERESTS, THE "PUBLIC INTEREST." AS HERRING OBSERVED IN THE MIDST OF THE CRISIS OF THE GREAT DEPRESSION, THE CLASH OF COMPETING INTEREST GROUPS DOES NOT NECESSARILY GUARENTEE THAT THE PUBLIC INTEREST, EVEN IF IT CAN BE DEFINED, WILL BE PROTECTED. THEREFORE, WHILE AMERICANS STRUGGLE WITH THE PROBLEMS OF ADMITTING NEW GROUPS INTO THE DECISION-MAKING PROCESS, THEY MUST ALSO GUARD AGAINST SACRIFICING THOSE INTERESTS WHICH WILL BEST SERVE THE NATION AS A WHOLE. THE LOCUS OF SOCIETAL DECISION-MAKING AUTHORITY WILL SHIFT FURTHER AWAY FROM LOCAL SCHOOL DISTRICT LEVELS TO STATE CAPITALS AND WASHINGTON, D.C.

Document	Abstract
Total Words 2377	Total Words 416
Data Elements 173	Data Elements 29
	Content .1676
	Length .1750
	DC 0.958

Figure 4.20 Computer-produced abstract of "Educational Decision Making"

FAMILIES CAN BE UNHEALTHY, FOR CHILDREN AND OTHER LIVING THINGS. #ARLENE SKOLNICK,PSYCHOLOGY TODAY 5(3), 18-22,104-106 AUG.1971.# THE MYTH OF THE FAMILY BLINDS US TO DANGERS OF OUR NORMAL CHILD-CARE PRACTICES. THERE IS VERY LITTLE TO PREVENT THE PARENT, IF HE IS SO INCLINED, FROM ACTING IN AN IRRATIONAL, UNFAIR, CRUEL, ABUSIVE OR SIMPLY NEGLECTFUL MANNER TOWARD HIS CHILD. INDEED THE PARENT IS LEGALLY EMPOWERED TO USE CORPURAL PUNISHMENT TO ENFORCE HIS RULES, NO MATTER HOW ARBITRARY THEY MAY APPEAR TO THE CHILD OR TO OTHERS. IF THE PARENT SHOULD KILL HIS CHILD IN THE COURSE OF ADMINISTERING A "DESERVED" BEATING, SOME STATES WOULD CONSIDER THE EVENT AN EXCUSABLE HOMICIDE. AT THIS POINT IT MAY SEEM THAT THE ARGUMENT HAS SLIPPED FROM A DISCUSSION OF PARENTS IN GENERAL TO AN EXTREME KIND OF BEHAVIOR--CHILD ABUSE AND MURDER. THIS BLURRING OF LINES IS INTENTIONAL. THE LITERATURE ON BATTERED CHILDREN YIELDS ONE MAJOR CONCLUSION: THERE IS NO CLEAR LINES OF DEMARCATION BETWEEN BATTERING PARENTS AND "NORMAL" ONES. NOTHING SETS THESE PARENTS OFF AS A GROUP IN TERMS OF SOCIAL CLASS, OCCUPATION, I.Q., URBAN-RURAL RESIDENCE, OR PSYCHOPATHOLOGY. ALL THAT RESEARCH HAS FO UND IS A PATTERN OF CHILD REARING THAT IS MERELY AN EXAGGERATION OF THE USUAL ONE. BATTERING PARENTS EXPECT STRICT OBEDIENCE FROM VERY YOUNG CHILDREN; THEY HAVE A CURIOUS SENSE OF RIGHTEOUSNESS; THEY FEEL THAT THEY ARE TEACHING THEIR CHILDREN NOT TO BE SPOILED AND DISRESPECTFUL.

Document	Abstract
Total Words 3434	Total Words 219
Data Elements 263	Data Elements 17
	Content .0646
	Length .0638
	DC 1.014

Figure 4.21 Computer-produced abstract of "Families Can be Unhealthy for Children and Other Living Things"

is designed to challenge popularly held views and the abstract reflects this tone. A person writing an abstract of this article would probably be tempted to make the abstract more objective in order to avoid the possibility of making unsupported statements. This tendency would not be desirable because the abstract would not convey the often inflammatory tone of the original article. Here the consistency and objectivity of the computer based abstracting system allows the subjectivity of the author to emerge. The data coefficient of 1.013 indicates that the abstract is an adequate replacement for the original in a condensed version.

The tenth sample abstract, shown in Figure 4.22 has the lowest data coefficient of the set of sample abstracts. The abstract tends to present disjoint ideas with no continuity between sentences. Intuitively, this abstract ranks lowest in the sample set, also. The original document, entitled "Automating Medical Records", appeared in the Delaware Medical Journal and was designed to acquaint physicians with recent developments in automation of hospital record keeping. The article presents a superficial survey of several developments and does not have a cohesive theme. The concluding sentence of the original, "It is important that all are aware of the developments in this field now.", indicates the intent of the article to publicize several new trends without actually reporting the details of these developments. The structure and style of the original document appears to be the significant limiting factor in the production of the abstract.

AUTOMATING MEDICAL RECORDS. #CAPT. CHARLES S. BURGER, DELAWARE MEDICAL JOURNAL 43(5), 127-129 (MAY,1971).# TWO DEVELOPMENTS IN COMPUTER TECHNOLOGY HAVE ALSO HASTENED DEVELOPMINT OF A COMPUTERIZED RECORD. THE FIRST IS THE TOUCH-SENSITIVE CATHODE RAY TUBE INPUT DEVICE WHICH ALLOWS VERY RAPID INPUT OF MEDICAL DATA. THE SECOND IS THE DEVELOPMENT OF "HIGH LEVEL" COMPUTER LANGUAGES TO ALTER THE MEDICAL CONTENT OF THE RECORD SYSTEM. THIS CAN BE LEARNED BY PHYSICIANS WITH GREAT EASE WITHOUT HAVING TO KNOW THE DETAILS OF COMPUTER OPERATIONS. UTILIZING THE FANTASTIC MEMORY AND RECALL CAPABILITY OF THE COMPUTER, WE CAN HOPEFULLY ELIMINATE ROTE MEMORY FROM OUR MEDICAL EDUCATION SYSTEM. THE CAPABILITY OF OBTAINING MULTIPLE PRINT-OUTS OF ALL OR PART OF THE MEDICAL RECORD COUPLED WITH PROBLEM-ORIENTED ORGANIZATION OF THE DATA WILL ALLOW AUDIT OF PHYSICIAN PERFORMANCE FOR THE PURPOSE OF DEFINING EDUCATIONAL NEEDS AND FOR MONITORING QUALITY OF PATIENT CARE.

Document	Abstract
Total Words 1174	Total Words 133
Data Elements 70	Data Elements 6
	Content .0857
	Length .1133
	DC .0.757

Figure 4.22 Computer-produced abstract of "Automating Medical Records"

The eleventh example, "Distribution of Attained Service in Time-Shared Systems", which is shown in Figure 4.23, is a technical article which presents a detailed discussion of queuing theory. The original article appeared in the Journal of Computer and System Sciences and presented several equations. In the published version the equations included greek letters, superscripts, subscripts and proofs for the equations. The input to the abstracting system was limited to the characters on the keypunch so the input of equations resulted in a loss of some of the data from the original article. The input of graphical material and special characters will certainly be a possibility with the development of advanced optical character recognition devices. The abstract has a data coefficient of 1.064.

The twelfth example has a data coefficient of 1.500 which is the highest of the sample set. This means that this abstract provides the best indication of content in the least nt of length with respect to its own document. The data coefficient does not measure the quality between a set of abstracts from different documents. The data coefficient can be used to indicate the best abstract among a set of abstracts that all represent the same original document.

The twelfth example, shown in Figure 4.24, comes from an unexpected source, Hot Rod Magazine and is entitled, "Auto Shop Series: Carburetion". The original article contains 2338 total words and the abstract contains only 25 words, or .011 of that number. The 25 words in the abstract appear in one sentence which was the last sentence in the original article. This sentence is really the autho-'s own summary

DISTRIBUTION OF ATTAINED SERVICE IN TIME-SHARED SYSTEMS, #L. KLEINROCK, JOURNAL OF COMPUTER AND SYSTEM SCIENCES:1,287-298(1967) # THE METHODS OF QUEUEING THEORY HAVE BEEN APPLIED TO A NUMBER OF SUCH MODELS TO OBTAIN VARIOUS MEASURES OF PERFORMANCE. IN THIS PAPER, WE CONSIDER A LARGE CLASS OF SUCH FEEDBACK QUEUEING SYSTEMS AND O BTAIN, FOR ALL OF THESE SYSTEMS, A RESULT WHICH DESCRIBES THE DISTRIBUTION OF ATTAINED SERVICE IN TERMS OF THE PREVIOUSLY SOLVED PERFORMANCE MEASURES. THE FOREGROUND-BACKGROUND SYSTEM IS ANOTHER EXAMPLE OF A MEMBER OF OUR CLASS. IN THIS SYSTEM, A NEW ARRIVAL JOINS THE FIRST QUEUE, OBTAINS A QUANTUM OF SERVICE, AND THEN, IF MORE SERVICE IS REQUIRE D, JOINS A SECOND QUEUE, ETC., JOINING THE NTH QUEUE ON HIS NTH VISIT TO THE SYSTEM OF QUEUES. THE SERVER ALWAYS GIVES SERVICE TO THE LOWEST NUMBERED QUEUE AND PROCEEDS TO THE NTH QUEUE ONLY IF THE N-1ST , ETC. QUEUES ARE EMPTY. THE SINGLE MOST SIGNIFICANT PERFORMANCE FACTOR OF ANY QUEUEING SYSTEM IS THE AVERAGE TIME THAT A CUSTOMER SPENDS WAITING IN QUEUES AS HE PASSES THROUGH THE SYSTEM. THE SERVICE FACILITY IN SUCH A CASE IS CONSTANTLY CYCLING AMONG DIFFERENT CUSTOMERS IN A CONTINUOUS WAY. IN A REAL SENSE, THEN, ALL CUSTOMERS PRESENT IN THE SYSTEM ARE USING A FRACTION OF THE SERVICE CAPACITY ON A FULL-TIME BASIS. INDEED, THE FRACTION OF THE MACHINE BEING USED BY A CUSTOMER FROM THE PTH PRIORITY GROUP AT SOME TIME T WHO HAS AN ATTAINED SERVICE IN THE INTERVAL(TAU,TAU+DT) IS MERELY G(TAU)/GAMMMAN(S)G(S) WHERE G(TAU)=LIM AS Q GOES TO.0.OG G(PN(Q)). SUCH AN OPERATING PROCEEDURE MAY BE REFERED TO AS A PROCESSOR-SHARED SYSTEM AND A DISCUSSION OF ITS BEHAVIOR MAY BE FOUND JACM 14. THE USEFULNESS OF THIS LIMIT OF PROCESSOR SHARING LIES IN ITS REPRESENTATION OF AN IDEALIZED SHARING OPERATION IN WHICH SWAP, TIME IS ASSUMED TO BE ZERO. THIS ASSUMPTION OF SWAP TIME IS AN IMPORTANT SIMPLIFICATION IN THESE MODELS. THE RESULTS THUS OBTAINED ARE IDEALIZED IN THE SENSE THAT NONZERO SWAP TIME CAN ONLY DEGRADE THE PERFORMANCE OF SUCH A SYSTEM. MODELS WITH NONZERO SWAP TIME HAVE BEEN CONSIDERED IN THE LITERATURE. THE RESULTS OF THIS PAPER GIVE GENERAL EXPRESSIONS FOR THE DISTRIBUTION OF ATTAINED SERVICE FOR ANY MEMBER OF A WIDE CLASS OF TIME-SHARED SYSTEMS, INCLUDING THSE WITH PRIORITY INPUTS. THE ANSWERS ARE GOOD FOR FINITE SERVICE QUANTA AS WELL AS FOR SERVICE QUANTA APPROACHING ZERO.

Document	Abstract
Total Words 2165	Total Words 385
Data Elements 148	Data Elements 28
	Content .1892
	Length .1778
	DC 1.064

Figure 4.23 Computer-produced abstract of "Distribution of Attained Service in Time-Shared Systems"

AUTO SHOP SERIES: CARBURETION. #DR. DEAN HILL, HOT ROD MAGAZINE 24(1D), 9D-93 (1971).# FOR THE CASUAL, BUT INTERESTED, "LET'S MAKE IT RUN BETTER" PERSON, CONSTANT MAINTENANCE, CARE AND ADJUSTMENT ARE NECESSARY FOR MAXIMUM RESULTS FROM ANY CARBURETION SYSTEM.

Document	Abstract
Total Words 2338	Total Words 25
Data Elements 187	Data Elements 3
	Content .0160
	Length .0107
	DC 1.500

Figure 4.24 Computer-produced abstract of "Auto Shop Series: Carburetion"

of the article and probably the best single sentence of the entire article to express the main ideas. The sentence reflects the style and informal tone of the entire article. It is not traditional to consider abstracts of articles from such sources as Hot Rod Magazine, but it is clearly possible to apply the abstracting system in diverse areas and achieve excellent results.

The thirteenth example, which is from an article which appeared in Fortune, is shown in Figure 4.25. In the article, "A Computer Version of How a City Works", the author, John F. Kain, writes a critical review of the work of Jay W. Forrester. The abstract reflects Kain's opinions which are positive on some aspects of Forrester's work and negative on others. The paragraph is coherent and provides a good representation of the original. It has a data coefficient of 1.202 for .173 of the content in .144 of the length of the original.

The final example is shown in Figure 4.26 and is an article from Scientific American. The article on "The Origins of Hypodermic Medication" provides a survey with anecdotes, quotations, and illustrations of the history of this medical practice. The original article contained 4297 total words and the abstract represents .061 of that length. The abstract is short enough to be included in a single paragraph but it does not provide a completely adequate reflection of the document's content. The style in which the original article was written seems to be the limiting factor in the application of the abstracting system.

A COMPUTER VERSION OF HOW A CITY WORKS. #JOHN F. KAIN, FORTUNE 80(6), 241-242 (1969).# FORRESTER, A PROFESSOR AT M.I.T.'S SLOAN SCHOOL OF MANAGEMENT, RELIES ON A COMPUTER MODEL HE DEVELOPED TO SIMULATE THE GROWTH, DECLINE, AND STAGNATION OF A HYPOTHETICAL CITY FROM BIRTH TO OLD AGE (250 YEARS). SUCH METHODS HAVE A GREAT DEAL OF POTENTIAL FOR THE ANALYSIS OF URBAN PROBLEMS AND HAVE ALREADY DEMONSTRATED THEIR VALUE IN A NUMBER OF SPECIFIC, THOUGH LIMITED APPLICATIONS. HOWEVER, THE DEVELOPMENT OF TRULY USEFUL AND TRUSTWORTHY URBAN SIMULATION MODELS REMAINS A DISTANT OBJECTIVE AND WILL REQUIRE MUCH GREATER RESOURCES THAN HAVE YET BEEN DEVOTED TO THE TASK. BEFORE ADEQUATE MODELS BECOME AVAILABLE, MANY INADEQUATE ONES WILL BE PUT FORWARD. FORRESTER'S MODEL IS A CONSPICUOUS EXAMPLE. IN HIS FIRST CHAPTER FORRESTER WARNS THE READER THAT CAUTION SHOULD BE EXERCISED IN APPLYING THE MODEL TO ACTUAL SITUATIONS. SUBSEQUENTLY, HOWEVER, HE EXPRESSES FEW RESERVATIONS ABOUT THE MODEL'S VALIDITY AND FREELY USES IT AS A BASIS FOR PRESCRIBING PUBLIC POLICY. THE INFLUENCE OF TAX RATES ON EMPLOYMENT AND POPULATION STRUCTURE IN FORRESTER'S CITY IS POWERFUL AND PERVASIVE. "MANAGERIAL-PROFESSIONAL" AND "LABOR" FAMILIES ARE ASSUMED TO BE REPELLED BY HIGH TAX RATES, WHEREAS THE UNDEREMPLOYED ARE INDIFFERENT TO THEM. HIGH TAX RATES, MOREOVER, DISCOURAGE THE FORMATION OF NEW ENTERPRISES AND ACCELERATE THE AGING OF EXISTING ONES. THERE ARE STILL OTHER ADVERSE EFFECTS: HIGH TAXES RETARD CONSTRUCTION OF BOTH PREMIUM AND WORKER HOUSING, WHICH IN TURN DISCOURAGES THE KINDS OF PEOPLE WHO LIVE IN THESE KINDS OF HOUSING FROM MOVING INTO THE CITY OR REMAINING THERE.

Document	Abstract
Total Words 1699	Total Words 244
Data Elements 139	Data Elements 24
	Content .1727
	Length .1436
	DC 1.202

Figure 4.25 Computer-produced abstract of "A Computer Version of How a City Works"

THE ORIGINS OF HYPODERMIC MEDICATION. #NORMAN HOWARD-JONES,SCIENTIFIC AMERICAN 224(1), 96-102 (JANUARY 1971).# THE AVAILABILITY OF COMPOUNDS THAT WERE HIGHLY ACTIVE IN MINUTE DOSES WAS A STIMULUS TO THE SEARCH FOR NEW METHODS OF ADMINISTERING THEM. LATER ANOTHER ARROW POISON, CURARE WAS TO BECOME A VALUABLE AID IN MODERN ANESTHESIOLOGY.) THE NOTION OF ADMINISTERING DRUGS THROUGH THE SKIN GAINED GROUND SLOWLY. IN 1859, AFTER WOOD'S METHOD HAD BECOME KNOWN, THE FRENCH PHYSICIAN LOUIS JULES BEHIER PUBLISHED AN ACCOUNT OF HIS OWN EXPERIENCE WITH HYPODERMIC MEDICATION. HE HAD USED A SYRINGE THAT DIFFERED FROM THE ONE EMPLOYED BY WOOD, AND HE CALLED IT THE "PRAVAZ APPARATUS." PRAVAZ HAD USED THIS TYPE OF SYRINGE TO INJECT CHEMICAL COAGULANTS INTO THE BLOOD VESSELS OF ANIMALS BUT NEVER FOR HYPODERMIC MEDICATION EITHER IN HIS PATIENTS OR IN ANIMALS. BEHIER'S USE OF PRAVAZ' NAME LED TO THE BELIEF THAT IT WAS HE WHO HAD INVENTED THE HYPODERMIC SYRINGE AND INTRODUCED HYPODERMIC MEDICATION INTO PRACTICE. IN THAT YEAR THE BRITISH PHYSICIAN J. M. CROMBIE, GIVING AS HIS JUSTIFICATION "THE DEARNESS AND DELICACY OF THE SYRINGE," ADVOCATED THE USE OF MORPHINE-IMPREGNATED SILK THREADS SLOWLY DRAWN THROUGH TWO PERFORATIONS IN THE SKIN BY MEANS OF A SUITABLE NEEDLE. THE NEW ALKALOIDS HAD BEEN DISCOVERED; IN THE CASE OF MEDICINES THUS INJECTED FOR GENERAL EFFECTS, HE CALLED THE METHOD 'HYPODERMIC,' TO DISTINGUISH IT FROM THE ENDERMIC, AND FROM THE LOCAL INJECTION OF WOOD.... THE HYPODERMIC DIFFERS FROM THE 'METHOD OF WOOD.' THE LATTER PLAN HAS FOR ITS OBJECT THE LOCAL TREATMENT OF A LOCAL AFFECTION." UNTIL THEN WOOD HAD NOT CONTESTED HUNTER'S CLAIM, BUT THIS STATEMENT GOADED HIM INTO A REPLY THAT WAS TO LEAD TO AN ACRIMONIOUS AND UNDIGNIFIED PUBLIC EXCHANGE OF CORRESPONDENCE. IT TOOK THAT LONG FOR THE COMPLICATIONS ATTENDING HYPODERMIC MEDICATION TO BE WIDE ACKNOWLEDGED BY THE MEDICAL PROFESSION.

Document		Abstract	
Total Words	4297	Total Words	264
Data Elements	349	Data Elements	19
		Content	.0544
		Length	.0614
		DC	0.886

Figure 4.26 Computer-produced abstract of "The Origins of Hypodermic Medication"

8.3. Summary of Results of the Experimental Application of the Evaluation Criterion

For each of the fourteen sample abstracts content, length, and data coefficient are tabulated in Table 4.3. The Content ranges from a low of .016 to a high of .237 with an average of .129. The Length ranges from a low of .011 to a high of .205 with an average of .123. The values of the data coefficient range from 0.757 to 1.500 with an average of 1.063. For the six documents that had fewer than 2000 total words, examples number 2, 3, 5, 6, 10 and 13, the average length of the abstract was .143 of the original and the average data coefficient was 1.071. For the four documents that had between 2000 and 2500 total words, examples number 4, 8, 11, and 12, the average length was .114 of the original and the average data coefficient was 1.045. For the four documents with more than 2500 words, examples number 1, 7, 9, and 14, the average length was .103 of the original and the average data coefficient was 1.006. These data show that the abstracting system produces relatively shorter abstracts for longer documents. They also show that the abstracts of longer documents are of poorer quality, although they are still above the minimum level of acceptability.

The average data coefficient for the sample abstracts produced by the abstracting system was 1.063. This indicates that the system produces abstracts which are acceptable, but not outstanding. The abstracts had low data coefficients because they were too long for the quantity of data retained. It would be possible to improve these abstracts by incorporating a procedure to edit the output from

Table 4.3 Summary of results of the examination of fourteen computer-produced abstracts

Example Number	Content	Length	Data Coefficient
1	.0813	.0981	.829
2	.1897	.2048	.926
3	.0938	.0861	1.088
4	.1019	.0907	1.124
5	.2374	.1638	1.450
6	.1402	.1472	.953
7	.2121	.1867	1.137
8	.1676	.1750	.958
9	.0646	.0638	1.014
10	.0857	.1133	.757
11	.1892	.1776	1.064
12	.0160	.0107	1.500
13	.1727	.1436	1.202
14	.0544	.0614	.886
Average	.1290	.1231	1.063

the abstracting system. The improvement should be based on a reduction in length without a corresponding reduction in content. The following chapter provides some techniques for the improvement of abstracts based on this conclusion.

References

1. Final Report on the Study for Automatic Abstracting, Thompson Ramo Wooldrige, Inc., Canoga Park, California, 1961 (PB 166 532).

2. J. E. Rush, R. Salvador, and A. Zamora, "Automatic Abstracting and Indexing. II. Production of Indicative Abstracts by Application of Contextual Inference and Syntactic Coherence Criteria," Journal of the American Society for Information Science 22 (4), 260-274 (1971).

3. A. DeLucia, "Index-Abstract Evaluation and Design," American Documentation 15 (2), 121-125 (1964).

4. H. P. Edmundson, "New Methods in Automatic Extracting," Journal of the Association for Computing Machinery 16 (2), 264-285 (1969).

5. D. Payne, J. Altman, and S. J. Munger, A Textual Abstracting Technique, Preliminary Development and Evaluation for Automatic Abstracting Evaluation, American Institute for Research, Pittsburgh, Pennsylvania, 1962 (AD 285 032).

6. D. Payne, Automatic Abstracting Evaluation Support, American Institute for Research, Pittsburgh, Pennsylvania, 1964 (AD 431 910U).

7. B. H. Weil, "Standards for Writing Abstracts," Journal of the American Society for Information Science 21 (5), 351-357 (1970).

8. G. J. Rath, A. Resnick, and T. R. Savage, "Comparisons of Four Types of Lexical Indicators of Content," American Documentation 12 (2), 126-130 (1961).

9. A. Resnick, "Relative Effectiveness of Document Titles and Abstracts for Determining Relevance of Documents", Science 134 (3438), 1004-1006 (1961).

10. G. J. Caras, "Indexing from Abstracts of Documents," Journal of Chemical Documentation 8 (1), 20-22 (1968).

11. B. C. Landry and J. E. Rush, "Toward a Theory of Indexing -- II", Journal of the American Society for Information Science 21 (5), 358-367 (1970).

12. B. C. Landry, A Theory of Indexing: Indexing Theory as a Model for Information Storage and Retrieval, Computer and Information Science Research Center, The Ohio State University, Columbus, Ohio, 1971 (OSU-CISRC-TR-71-13).

13. M. C. Yovits and R. L. Ernst, "Generalized Information Systems," in A. Kent, O. E. Taulbee, J. Belzer, and G. D. Goldstein (eds.), Electronic Handling of Information, Thompson Book Company, Washington, D.C., 1967, 279-290.

14. M. C. Yovits and R. L. Ernst, "Generalized Information Systems: Consequences for Information Transfer," in H. B. Pepinsky (ed.), People and Information, Pergamon Press, Elmsford, New York, 1970.

15. H. Borko and S. Chatman, "Criteria for Acceptable Abstracts: A Survey of Abstracters' Instructions," American Documentation 14 (2), 149-160 (1963).

16. P. Beckmann, The Structure of Language--A New Approach, The Golem Press, Boulder, Colorado, 1972.

17. C. A. Montgomery, "Linguistics and Information Science," Journal of the American Society for Information Science 23 (3), 195-219 (1972).

18. C. E. Young, Design and Implementation of Language Analysis Procedures with Application to Automatic Indexing, Ph.D. dissertation, The Ohio State University, Columbus, Ohio, in preparation.

19. K. Hirayama, "The Length of an Abstract and Amount of Information", Journal of Chemical Documentation 15 (2), 121-125 (1964).

CHAPTER V. IMPROVEMENT OF THE ABSTRACTING SYSTEM

1. Improvement of the Quality of Abstracts through Sentence Modification[1]

1.1. Rationale for Sentence Modification

Most existing automatic abstracting systems only select sentences from the original document to form an extract. But sentences which are well-suited to the document in their original context, may not be well-suited for inclusion in the abstract because of the new context in which they are found. Frequently, the set of selected sentences provides only a set of disjoint ideas in unrelated sentences. Wyllys considers this property of extracts to be a major disadvantage.

> The most serious disadvantage of current computer produced abstracts is that they consist of individual sentences of the original text, extracted according to one or more criteria. Not only do the extraction criteria require further research, but the resulting set of individual sentences presents problems of disjointness, incompleteness, redundancy, and the like. The ultimate goal of research in automatic abstracting is to enable a computer program to "read" a document and "write" an abstract of it in conventional prose style, but the path to this goal is full of unconquered obstacles. (1)

Conventional prose style dictates that appropriate connnectives and transitions between ideas be provided so that the reader is not required

---

[1] The work reported in Section 1 of this chapter was carried out with C. E. Young. A paper, entitled "Improvement of Automatic Abstracts by the Use of Structural Analysis," reports the results of this research and has been submitted for publication in the Journal of the American Society for Information Science. I thank Miss Young for permitting me to include this material in my dissertation.

to guess at what these should be. An abstract may be difficult to understand if the ideas it expresses are not presented and connected in a logical way. As Needleman puts it, unity of a document requires, primarily, a judicious assembling or blending of ideas and details; and coherence is achieved when these details or ideas are so arranged and so worded that there is a clear, continuous, logical progression of thought from sentence to sentence (2).

In order to provide a logical progression of thought within an abstract, sentences should be related to the preceding and following sentences, or a clear transition of thought should be provided. The relationship between sentences may be expressed by the repetition of keywords and by the use of connectives to bridge the gaps in thought between sentences. Not every statement demands a transitional link, but a statement may often need to be connected in some way with those in juxtaposition with it. This is especially so when sentences appear to be loosely connected, yet have real sequences of ideas that should be brought into clear relation by the use of stylistically strategic transitional devices.

The primary goal in the development of ADAM was to construct an algorithm which would delete all sentences from the original document that were not worthy of inclusion in the abstract, while retaining those sentences that did not satisfy the deletion criteria. In ADAM, each sentence of the original document is first evaluated individually. If the sentence is considered worthy of inclusion in the abstract, but requires an antecedent, the three sentences preceding the selected

sentence are examined to determine whether they should also be included in the abstract. The requirement of an antecedent is established if the selected sentence contains certain function words or phrases, such as "for that reason", "hence", "in that case", "their", "these", "they", "this", and "those". Such a criterion provides a good method of finding sentences which should be included in an abstract because of their relationship to other sentences selected for inclusion through application of the principal selection criteria. There are, however, occasions when this technique is not effective. Those include cases in which the related sentences are separated by more than three sentences, or in which intersentence relationships are indicated by devices other than the function words or phrases normally employed. These occasions occur frequently enough to warrant the development of other methods for the detection of intersentence references. Three alternative situations may arise in attempts to handle intersentence references. If a sentence is judged abstract worthy and it contains an indication of intersentence reference, then

a. the related sentences must be found in the document and included in the abstract; or

b. if the related sentences cannot be found, the selected sentence must be rewritten to make it stand alone; or

c. if neither of these cases applies, delete the selected sentence.

The problems associated with the location of intersentence referents and with the rewriting of sentences of the original document, and

possible approaches to their solution, are considered in the next section.

1.2. The Structural-Analytic Approach to Sentence Modification

In this research procedures have been developed for the modification of sentences initially selected by ADAM (all processes to be described are performed on the abstract, not the original document). These procedures include

a. a method of identification of words and phrases of potential importance to the abstract;

b. a means of retention or deletion of certain clauses or sentences;

c. a method for the revision of and, in some instances, the creation of sentences.

These procedures have been developed using the notions of structural analysis (3) and employing a number of programs that facilitate this analysis.

The structural approach to linguistic analysis employed in this research is based upon the notions of Fries (3), who criticized the traditional approach to grammatical analysis for its lack of consistency of definition of grammatical classes. To Fries, grammatical classes should be defined on the basis of usage, rather than on the basis of "meaning". Thus, four main classes of words were defined based upon the positions the words could occupy in one or more simple frame sentences. For instance, the sentence

The concert was good.

serves as a frame such that any word that can replace 'concert' in the frame is a member of Class 1, while any word that can replace 'was' is a member of Class 2. Similarly, words that can replace 'good' in the sentence frame are members of Class 3. Class 4 words are those that can replace 'there' in the sentence frame

The team went there.

These four classes of words correspond roughly to the traditional classes 'noun', 'verb', 'adjective' and 'adverb', but the classes are defined strictly on the basis of structure.

In addition to the four classes mentioned above, Fries defined a fifth class of words which is composed of fifteen groups of specific elements. Some of these groups include prepositions, conjunctions, pronouns, auxiliary verbs and determiners (the articles, among others). This fifth class, called function words, differs from the other four in that its members

a. serve as relators between groups of words in the other four classes, as well as structural signals within the sentence;
b. have no meaning ascribed to themselves, but must be known as specific items;
c. constitute a small set of elements (Fries identifies 154) which generally account for between 45% and 50% of all word occurrences in a given body of text (4).

Using the ideas of Fries as a basis, a program has been devised that identifies the class to which each word in an English sentence belongs. This program, called MYRA, has been described elsewhere (5). It served as one of the principal tools of the present work, and its use will become obvious shortly.

Two additional language analysis programs which have been used in this study include a phrase analysis program and a clause analysis program. The phrase program:

1. identifies and isolates noun, verb, preposition and adjective phrases;
2. identifies the head word of each phrase;[2]
3. identifies the subject and object of each sentence.[3]

The clause program identifies and isolates each clause within a sentence. The operation of the three programs mentioned above can be illustrated by means of a simple example. The sentence

The thief ran from the police.

contains the three function words 'the', 'from' and 'the'. The first 'the' signals the fact that either an adjective or a noun must follow. The preposition 'from' introduces a noun phrase, and the second 'the'

---

[2] The head word of a phrase is that word which all other words in the phrase modify.

[3] The program referred to contains an implementation of case grammar (5). The major case assignments are comparable to the traditional notions of subject and object.

causes 'police' to be identified as a noun. Assuming that 'thief' were identified as an adjective initially, and that 'ran' were identified as a noun, MYRA would note that the string of words contained no verb, so that an attempt to reassign some of the words would be made. In this example, the only logical place for a verb to occur is before the preposition. Hence, 'ran' would be reassigned as a verb and 'thief' would be classed as a noun. The results of this analysis form the input to the phrase identification program.

The phrase identification program interprets the determiner 'the' as initiating a noun phrase. The program continues to scan to the right in the sentence until a noun is located. The phrase is thus delimited, and the head word (the noun) is also identified. Other types of phrases are similarly identified and delimited.

The clause identification program depends, for its proper functioning, upon the grouping of the phrases that follow the verb phrase(s), and upon the identification of coordinate conjunctions, subordinate conjunctions, relative pronouns and words such as 'that' and 'there.' This program is based upon the definition of a clause as a sequence of words that contains one and only predicate. This definition is due to Cook (6). Consider, for example, the sentence

The company admitted that they were wrong.

Two verb phrases would be recognized by the phrase identification program, namely, 'admitted' and 'were'. The structural signal that differentiates the two clauses in this sentence is the word 'that'.

Therefore, the two clauses are

The company admitted
(that) they were wrong.

The three programs described briefly above are the basic analytical tools used in this research. A complete description of the programs is given in (5, 7). Their application will be made evident in the next section.

1.3. Rules for Sentence Modification

Five specific rules for effecting modifications on sentences of an abstract have been developed, based upon data provided by the structural-analytic programs mentioned in the preceding section. These rules are

1. Combination of sentences by means of a coordinate conjunction.
2. Combination of sentences by means of a subordinate conjunction.
3. Modification of sentences by means of a graphical reference transformation.
4. Creation of sentences by means of a reference tabulation.
5. Revision or deletion of sentences for context modification.

These rules will be described in detail in the following sections. The symbology employed in these descriptions is defined in Figure 5.1.

1.3.1. Combination of Sentences

The primary criterion used to determine whether sentences might be suitably combined is that of parallelism. For two sentences to

NONP	Noun phrase
VRBP	Verb phrase
PRPP	Prepositional phrase
PRNP	Pronoun phrase
...	Continuation
...P*	Modified phrase
$...P_1$	Phrase required for the application of a rule
$...P_a$	Phrase which is not necessary for specific consideration in the application of a rule

Figure 5.1 Symbology employed in the description of rules for sentence modification.

read smoothly in combination, there must be parallelism of structure and continuity of thought. Two sentences are said to have parallel structures if they have the same ordering of dependent and independent clauses and a similar ordering of phrases (by type). The strict parallelism in the ordering of sentence elements is relaxed in the case of prepositional phrases, wherein both the number and order of the phrases may differ among the sentences combined.

For a determination of continuity of thought, it has been found adequate to test for identical main verbs or for identical subjects in the two sentences under consideration. Such a test does not, of course, in the case of non-identical construction, mean that continuity of thought does not exist, but such identity insures that the sentences will be appropriately combined.

A second criterion that can be used to determine whether two sentences should be combined is that of sentence complexity. Sentences that are quite long and of complex construction (such as the last sentence of the previous paragraph) are more difficult to read than shorter, simpler sentences and they are generally undesirable in an abstract. Thus two sentences would not be combined if one or both contain too many clauses. I have stipulated that sentences may not be combined if they have more than one independent and more than one dependent clause.

1.3.1.1. Combination of Sentences by Means of a Coordinate Conjunction

Three alternative rules have been developed for combining sentences by means of a coordinate conjunction. These rules can be applied to

any pair of sentences which pass tests for similarity of structure and for similarity in their subject and verb phrases. Every pair of sentences must satisfy the following structural rules:

1. Determine that each sentence under consideration has only one independent clause and at most one dependent clause.
2. Determine that the sentences are of parallel construction. The order and type of clause must be the same in both sentences and there must be similarity of phrase structure.

Each rule has a set of associated similarity tests. For each rule, the following tests must be made:

Rule 1.a - identical main verb phrases

Rule 1.b - identical subject noun phrases

Rule 1.c - identical subject noun phrases and identical main verb phrases

Rule 1.d - identical head word of the subject noun phrases and identical main verb phrases.

The most general of the rules is that which requires identity of both main verbs. Auxiliaries and adverbs contained in the verb phrase may vary from sentence to sentence; only the main verbs must be the same. For instance, the verb phrases 'can provide' and 'cannot generally provide' satisfy the criterion of identical main verb and so would permit the combining of the sentences. To clarify the issues so far discussed, consider the following sentences.

'Thus, the profession of medicine provides services which utilize its knowledge of anatomy, physiology, neurology, pathology, and such areas.'[4]

'The legal profession provides services which utilize its special knowledge of jurisprudence.'

Once the classes to which the words in the sentences belong have been determined, the clause program would identify two clauses in each sentence. For the first sentence, these are

'Thus, the profession of medicine provides services'

and

'which utilize its knowledge of anatomy, physiology, neurology, pathology, and such areas.'

The second clause is identified as a dependent clause since it begins with the relative pronoun 'which'. An identical clause construction is found in the second sentence. Application of the phrase analysis program to these sentences yields a determination that the phrase structure of the sentences is similar. The complete analysis of these two sentences is shown in Figure 5.2. Since the main verb in the two sentences is the same, the sentences meet all the criteria for combination. Using the conjunction 'and' for this purpose, we obtain the new sentence

'Thus, the profession of medicine provides services which utilize its knowledge of anatomy, physiology, neurology, pathology, and such areas and the legal profession provides services which utilize its sepcial knowledge of jurisprudence.'

---

[4] All sample sentences are taken from abstracts produced by ADAM.

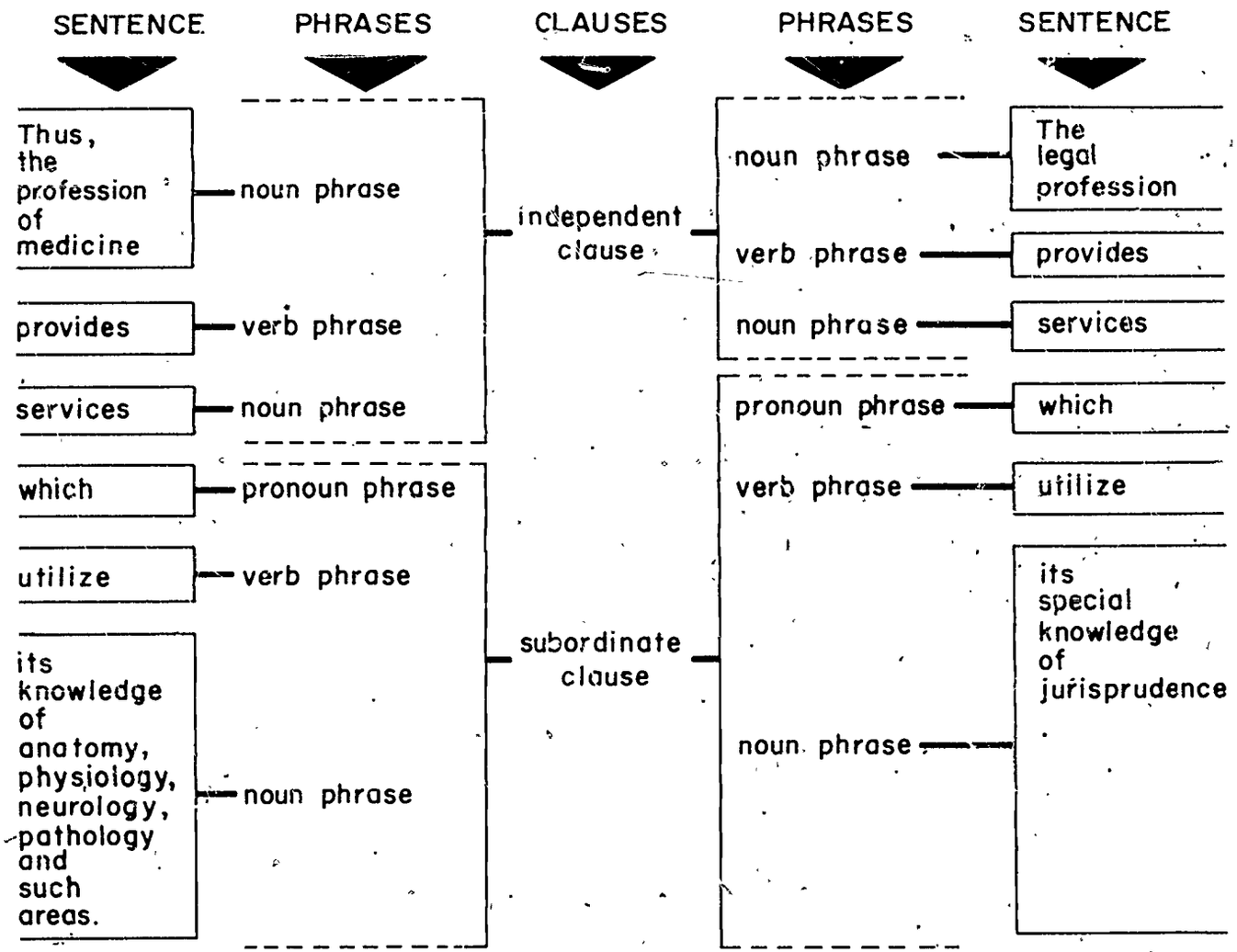


Figure 5.2 Structural analysis of two sample sentences which may be combined with a coordinate conjunction.

The rule for combining sentences that have identical main verbs may be more generally expressed as follows.

$$S_1 = NONP_a\ VRBP_1\ NONP_b \mid PRN_c\ VRBP_d\ NONP_e$$
$$+$$
$$S_2 = NONP_f\ VRBP_2\ NONP_g \mid PRN_h\ VRBP_i\ NONP_j$$
$$\rightarrow$$
$$S_3 = NONP_a\ VRBP_1\ NONP_b\ PRN_c\ VRBP_d\ NONP_e\ \text{'AND'}$$
$$NONP_f\ VRBP_2\ NONP_g\ PRN_h\ VRBP_i\ NONP_j$$

A second rule for combining sentences, one slightly less general than the above rule, requires the presence of identical subjects in the sentences under consideration. If the noun phrases that precede the main verb of a pair of sentences match word for word, then the noun phrase of the second sentence is deleted and the remaining sentence fragment is combined with the first sentence by means of a coordinate conjunction. In the two sentences

'The system exceeded the capacity of its present auxiliary equipment.'

and

'The system was modified for further testing.'

the subject noun phrases are found to be identical, so the subject of the second sentence is deleted and the remainder combined with the first to yield the sentence,

'The system exceeded the capacity of its present auxiliary equipment and was modified for further testing.'

Expressed symbolically, this process is

$$S_1 = NONP_1\ VRBP_a\ NONP_b$$
$$+$$
$$S_2 = NONP_1\ VRBP_c\ NONP_d$$
$$+$$
$$S_3 = NONP_1\ VRBP_a\ NONP_b\ \text{'AND'}\ VRBP_c\ NONP_d$$

The third rule for combining sentences is the most specific of the three because it requires that two sentences have identical subject noun-phrases and identical verb phrases if they are to be combined. The following sentences satisfy this criterion.

'The experiment resulted in a modification of the original hypothesis.'

'The experiment resulted in a change of our basic approach.'

Both of these sentences pass the tests for parallelism in clause and phrase structure, and both contain subject and verb phrases that match word for word. Thus, both the subject and verb phrases are deleted from the second sentence and the remaining fragment is combined with the first sentence by use of a coordinating conjunction, yielding the sentence

'The experiment resulted in a modification of the original hypothesis and in a change of our basic approach.'

This sentence combination process can be expressed symbolically as follows.

$$S_1 = NONP_1\ VRBP_1\ PRPP_a$$
$$+$$
$$S_2 = NONP_1\ VRBP_1\ PRPP_b$$
$$+$$
$$S_3 = NONP_1\ VRBP_1\ \ldots\ \text{'AND'}\ \ldots$$

A variation of the third rule is the fourth rule which can be employed when a single noun phrase as subject is present in both sentences under consideration. This rule requires only that the head word of each subject noun phrase be identical, in addition to identity of verb phrases, in order for the sentences to be combined. In this case, however, both the subject modifiers (except determiners) and the sentence predicates must be conjoined by a conjunction. The word "respectively" is added at the end of the sentence to maintain the proper logical ordering. The following sentences serve for illustration. Application of the rule just described to the sentences

'Individual manufacturers offer ALGOL, BASIC, and FOCAL compilers.'

'Cost manufacturers offer programming support on an individually negotiated contract basis.'

yields the sentence

'Individual and cost manufacturers offer ALGOL, BASIC, and FOCAL compilers, and programming support on an individually negotiated contract basis, respectively.'

In applying this rule, the verb must be converted to plural form if it is not already plural. If the verb phrase is initiated by an auxiliary verb (a form of the verb 'to be' or 'to have') the pluralization process is straightforward. The number of the verb is determined by comparing the auxiliary with a dictionary containing the inflected forms of these verbs. If the auxiliary in the sentence matches one in the dictionary that is of singular form, the verb phrase of the sentence is appropriately pluralized.

If an auxiliary verb does not initiate the verb phrase, and provided the subject of the sentence is in the third person, the pluralization process is based upon the following observations rather than on dictionary look-up. If a verb ends in 'ies' and contains fewer than five characters, it is pluralized by dropping the 's'. Thus verbs like 'lies' a. 'dies' would be appropriately transformed into their plural forms. If a verb, ending in 'ies', is five or more characters in length, then the 'ies' is changed to 'y' to yield the plural. 'Tries' and 'flies' would thus be satisfactorily pluralized.

If the verb ends in 'oes' or 'ses', as in 'goes' and 'stresses', then the 'es' is dropped to form the plural. Finally, if none of these rules applies and the verb ends in 's', then the letter 's' is deleted to form the plural.

If the verb is singular, then the head word of the subject noun phrase must be pluralized as well as the verb phrase. The pluralization of nouns is not so easy a task as is the pluralization of verbs. Nevertheless, programs for the purpose have been developed

which produce satisfactory results; I have chosen to use the procedures developed by Petrarca and Lay (8).

After the nouns and verbs have been pluralized, a final check for the indefinite articles 'a' and 'an' is made. If these articles are found then they are deleted. The sentences are thus made ready for combination, a process depicted formally as follows (where asterisks denote sentence elements potentially modified by the pluralization process or by deletion).

S_1 = $ADJP_1$ $NONP_1$ $VRBP_1$ $NONP_a$

\+

S_2 = $ADJP_2$ $NONP_1$ $VRBP_1$ $NONP_b$

\+

S_3 = $ADJP_1$* 'AND' $ADJP_2$* $NONP_1$* $VRBP_1$* $NONP_a$ 'AND' $NONP_b$

',respectively.'

1.3.1.2. Combination of Sentences by Means of a Subordinate Conjunction

A second general way of combining sentences is to find two whose structure is such that one sentence can be made a subordinate clause of the other. Procedures have been developed which search the abstract for sentences that can be thus combined. As in the case of sentence combinations effected through use of a coordinate conjunction, the application of this procedure requires that the candidate sentences contain only one independent and one dependent clause. Furthermore, the dependent clause must not begin with 'who' or 'which'.

This rule combines two sentences into a sentence with an independent clause and a subordinate clause. In order to determine

if a sentence can be made a subordinate clause, the subject of the sentence is compared with every other noun phrase in the abstract, (including noun phrases that are the object of a preposition.) If a match is found, then the sentence in which the noun phrase occurs becomes the independent clause. As an illustration of this rule, consider the following sentences.

'A set of consecutive storage locations is called a memory block.'

'A memory block is labelled by a single word called a codeword.'

The sentences are formally represented as:

$S_1 = NONP_a\ VRBP_b\ NONP_1$

$S_2 = NONP_1\ VRBP_c\ PRPP_d\ VRBP_e \ldots NONP_f$

The identical phrase in the sentences is $NONP_2$ (i.e., 'a memory block'). Sentence 2 will become a subordinate clause of sentence 1 because $NONP_2$ is the subject of this sentence. $NONP_2$ is deleted from sentence 2, and is replaced either by 'who' or by 'which'. The relative pronoun 'who' is associated only with nouns that have human attributes. Thus a personification test is made. This is done by checking the inflectional ending of the noun. If the word ends in 'ist(s)' or 'ian(s)', and if the word has more than 5 letters, as in scientist, physicians and librarian, then the noun is flagged with the human attribute, and the relative pronoun 'who' is added. If the noun is not flagged, the

relative pronoun 'which' is used. This rule is not adequate to detect all nouns that have human attributes, for example, names of individuals, but it provides an adequate means of deciding between 'which' and 'who' in most cases.

The new sentence has the form:

$$S_3 = NONP_a\ VRBP_b\ NONP_1\ ,\ which'\ VRPB_c\ PRPP_d\ VRBP_e\ NONP_f$$

For the sample sentences given above, the new sentence is:

'A set of consecutive storage locations is called a memory block, which is labelled by a single word called a codeword.'

The rules for combination of sentences by means of a coordinate or subordinate conjunction are summarized in Figure 5.3.

1.3.2. The Graphical Reference Rule

It is not desirable in an abstract, to refer to a specific graph, figure or table contained in the document. If, however, such references do find their way into the abstract, they can be readily identified and replaced by a general descriptive reference. Specific references may be identified through the use of a small dictionary of words signalling such references and by applying certain contextual tests. The dictionary includes such words as table, tab., figure, fig., graph, etc. And the contextual tests include the identification, following one of these words, of a) a string of digits, b) a single character, or c) a Roman numeral. If a contextual test is satisfied, the number or character following the dictionary word is deleted and

1. Combination of sentences by means of a coordinate conjunction

   a. Identical main verb phrases

   S_1 = $NONP_a$ $VRBP_1$ $NONP_b$ PRN_c $VRBP_d$ $NONP_e$
   +
   S_2 = $NONP_f$ $VRBP_2$ $NONP_g$ PRN_h $VRBP_i$ $NONP_j$
   ↓
   S_3 = $NONP_a$ $VRBP_1$ $NONP_b$ PRN_c $VRBP_d$ $NONP_e$ 'AND'
   $NONP_f$ $VRBP_2$ $NONP_g$ PRN_h $VRBP_i$ $NONP_j$

   b. Identical subject noun phrases

   S_1 = $NONP_1$ $VRBP_a$ $NONP_b$
   +
   S_2 = $NONP_2$ $VRBP_c$ $NONP_d$
   ↓
   S_3 = $NONP_1$ $VRBP_a$ $NONP_b$ 'AND' $VRBP_c$ $NONP_d$

   c. Identical subject noun phrases and identical main verb phrases

   S_1 = $NONP_1$ $VRBP_1$ $PRPP_a$
   +
   S_2 = $NONP_1$ $VRBP_1$ $PRPP_b$
   ↓
   S_3 = $NONP_1$ $VRBP_1$ ... 'AND' ...

   d. Identical head word of the subject noun phrases and identical main verb phrases

   S_1 = $ADJP_1$ $NONP_1$ $VRBP_1$ $NONP_a$
   +
   S_2 = $ADJP_2$ $NONP_1$ $VRBP_1$ $NONP_b$
   ↓
   S_3 = $ADJP_1$* 'AND' $ADJP_2$* $NONP_1$* $NONP_1$* $NONP_a$ 'AND' $NONP_b$',
   respectively'

2. Combination of Sentences by Means of a Subordinate Conjunction

   S_1 = $NONP_a$ $VRBP_b$ $NONP_1$
   +
   S_2 = $NONP_1$ $VRBP_c$ $PRPP_d$ $VRBP_e$ ... $NONP_f$
   ↓
   S_3 = $NONP_a$ $VRBP_b$ $NONP_1$ ', which' $VRBP_c$ $PRPP_d$ $VRBP_e$ $NONP_f$

Figure 5.3 Summary of rules for combination of sentences by means of a coordinate or subordinate conjunction

the initial word of the phrase is identified. If the initial word is 'the', this word is deleted. Finally, the article 'a' or 'an' is inserted (if not already present) at the beginning of the phrase.

Sentences such as:

S_1 'Table 2 presents nine areas of endeavor and their associated disciplines.'

S_2 'Figure 2 presents graphically the general model of information transfer.'

are modified to a general reference:

S_1 'A table presents nine areas of endeavor and their associated disciplines.'

S_2 'A figure presents graphically the general model of information transfer.'

by application of this rule.

1.3.3. Reference Tabulation

Literature references contained in a document are not included in an abstract, yet the number and kind of reference may indicate something of the strength of the paper. A means has therefore been devised of tabulating the references in a paper, based initially upon the format used by the Journal of the American Society for Information Science, and of presenting this data in a sentence of the abstract. The procedure developed can be described as follows.

The heading designated by 'References' is identified in the text. Next set N = 1, and search for the character string 'bN.b', where b indicates a blank. A variable REFERENCE_FLAG is set to zero. After 'b1.b' is located, REFERENCE_FLAG is set to 1, N is incremented by 1,

and the next consecutive number in the string 'b2.b' is sought. This process is continued until the text is exhausted or until another heading 'Appendix' is encountered. At this time, N is decremented by one. The REFERENCE_FLAG is checked and if it is equal to zero, the sentence 'No references are given.' is appended to the abstract. If REFERENCE_FLAG is non-zero, the sentence 'N references are given.' is appended to the abstract. This procedure is specific to journals that have this format for references, but modifications of this basic technique could be added to the system to allow for varying formats.

1.3.4. Context Modification

Sentences sometimes appear in an abstract which refer to sentences that appear in the original document but not in the abstract. Such sentences have been found generally to contain phrases of ordinal designation such as 'the second ...' or 'the first ...' before the main verb of the sentence. This is the case, for example, in the following sentences.

'The second mechanism is structural change: ...'

'The second is that reactions of oxygen atoms in the low temperature region tend to be more stereoscopic with trans- than with cis-olefins.'

'The first is the H12 developed by Honeywell.'

A sentence that contains an ordinal number, n, in the first phrase requires at least n - 1 antecedent sentences, the first of which indicates the points enumerated. If a sentence exhibits the fault

of referring to a sentence not in the abstract, the required antecedents could be searched for in the original document and, when found, added to the abstract. The time and effort involved in such a search would not, however, generally be acceptable. Hence, the following procedure has been developed for handling sentences of an abstract whose leading phrase(s) contain words which demand an antecedent.

If a sentence of an abstract contains a leading phrase which has an ordinal number as a component, the abstract is searched backward from this sentence to find sentences that contain lower ordinal numbers or else an appropriate cardinal number. For example, to complete the antecedent relationships for either of the first two sample sentences above, sentences must be found earlier in the abstract which contain the phrase 'the first ...' and the adjective 'two', respectively. If the required antecedent sentences are not found, then the sentence is handled in one of the following ways. If the ordinal number serves as an adjective in the sentence, then it can be deleted and the determiner 'the' is changed to 'a' or 'an', if necessary. The first example sentence above would become

'A mechanism is structural change: ...'

If, on the other hand, the ordinal number serves as a noun in the sentence in question, then the sentence is either deleted, as in the case of the third example sentence, above, or else, if the construction of the sentence follows the pattern

'The (ordinal number) is that ...'

the portion of the sentence up to and including 'that' is deleted and the rest of the sentence is allowed to remain. Thus, the second example sentence above would become

> 'Reactions of oxygen atoms in the low termperature region tend to be more stereoscopic with trans- than with cis-olefins.'

1.4. Quality Improvements Viewed According to the Evaluation Criterion

I have described several methods for improving the readability of abstracts produced by the abstracting system. These methods are designed to decrease the number of words needed to express a concept and to make the resulting abstracts more readable and coherent. These methods can be viewed in light of the evaluation criterion to study their effect on the quality of the final product. In this section I will present three examples of the results of the application of the rules for sentence modification.

The first example is the abstract of a chemistry article, "Addition of Oxygen Atoms to Olefins at Low Temperatures" which appeared in the Journal of Physical Chemistry. The abstract of this article, before any modification, appears in Figure 5.4. This abstract contains .158 of the content of the original in .167 of the length and has a data coefficient of 0.942. This abstract could be edited to reduce its length and improve its data concentration. Application of the sentence modification rules to this abstract results in the abstract shown in

ADDITION OF OXYGEN TO OLEFINS AT LOW TEMPERATURE IV REARRANGEMENTS. #R. KLEIN AND M.D. SCHEER, JOURNAL OF PHYSICAL CHEMISTRY 74(3), 613-616(1970). # A CONSIDERATION OF THE OXYGEN ATOM ADDITION TO CIS- AND TRANS-2-BUTENE IN THE TEMPERATURE REGION 77 TO 113 K LED TO THE FORMULATION OF A NEW TRANSITION INTERMEDIATE. IN THIS INTERMEDIATE, THE OXYGEN ATOM IS REPRESENTED AS BOUND IN A LOOSE, THREE-MEMBERED RING WITH, AND IN THE PLANE OF, THE OLEFINIC STRUCTURE OF THE REACTANT. OBSERVATIONS ON 2-BUTENES HAVE BEEN EXTENDED TO SEVERAL MORE STRAIGHT-CHAIN, INTERNAL OLEFINS IN THE LOW-TEMPERATURE REGION. COMPARISON OF THE TRANS-EPOXIDE TO KETONE RATIONS FROM THE CIS- VS THE TRANS-OLE FIN WITH INCREASING SIZE OF THE OLEFIN INDICATES THAT THESE RATIOS DIVERGE. THE SECOND IS THAT REACTION OF OXYGEN ATOMS IN THE LOW-TEMPERATURE REGION TENDS TO BE MORE STEROSCOPIC WITH TRANS- THAN WITH CIS-OLEFINS. CARBONYL COMPOUNDS CONSTITUTE A SIZEABLE FRACTION OF THE PRODUCTS OF THE OXYGEN ATOM ADDITION TO OLEFINS IN THE LOW-TEMPERATURE REGION AND, AS HAS BEEN NOTED, AN INTRAMOLECULAR GROUP MIGRATION IS REQUIRES FOR CARBONYL FORMATION. THE PRINCIPAL CARBONYL PRODUCT IN THE TRANS-2-BUTENE REACTION AT 90 K IS 2-BUTANONE. THE FORMATION OF THIS KETONE REQUIRES THE MIGRATION OF H. COMPARED TO THE MIGRATION OF THE METHYL GROUP, THAT OF H IS SLIGHTLY FAVORED. CIS-2-BUTENE IS NOT USEFUL FOR THE COMPARISON, AS BOTH OF THE HYDROGEN ATOMS ATTTACHED TO THE OLEFINIC CARBON PAIR ARE SUPPRESSED THROUGH INTERACTION WITH OXYGEN IN THE COMPLEX. THE RELATIVE QUANTITIES OF 2-BUTANONE TO ISOBUTYRALDEHYDE IS TAKEN AS A MEASURE OF THE RATIO OF MIGRATION OF THE HYDROGEN ATOM TO THE METHYL GROUP. REACTIONS WERE EFFECTED AT 90 K IN THE APPARTATUS ROUTINELY USED FOR THIS PURPOSE. THE OLEFINS WERE DILUTED 10 TO 1 WITH PROPANE. THE EXPOSURE TIME OF OXYGEN ATOMS WAS 5 MINUTES, AND ABOUT 1% OF THE OLEFIN WAS REACTED. THE PRODUCTS WERE DETERMINED AT 135 AND A HELIUM FLOW OF 100 CC/MINUTE. THE CIS AND TRANS ISOMERS OF 3,4-EPOXY-3,4-DIMETHYLHEXANE WERE NOT SEPARABLE. LOCALIZTION OF THE OXYGEN ATOM IN THE TRANSITION COMPLEX PRECEDING ALKYL GROUP REARRANGEMENT IS NOT IN ACCORD WITH THE EXPERIMENTAL RESULTS. AT 90 K, THE RATIO OF ADDITION TO C-2 IS 16 TIMES THAT TO C-3. FOR MEP, ADDITION OF THE OXYGEN ATOM TO THAT CARBON ATOM OF THE DOUBLE BOND TO WHICH THE TWO METHYL GROUPS ARE ATTACHED WOULD BE EXPECTED TO BE FAVORED.

Document	Abstract
Total Words 2201	Total Words 368
Data Elements 146	Data Elements 23
	Content .1575
	Length .1672
	DC 0.942

Figure 5.4 Computer-produced abstract of "Addition of Oxygen Atoms to Olefins at Low Temperatures"

Figure 5.5. The only rule that was applicable in this case was the rule for context modification. The sentence,

> 'The second is that reaction of oxygen atoms in the low-temperature region tends to be more stereoscopic with trans- than with cis-olifins.'

appears in the abstract but the required antecedent sentences are not contained in the abstract. The sentences which preceded this sentence did not contain the phrase 'the first' and the adjective 'two'. Since the ordinal number, in this case, 'second', appears in the pattern

> 'The (ordinal number) is that ...'

the portion of the sentence up to and including 'that' is deleted and the rest of sentence is allowed to remain. Thus the sentence becomes

> 'Reaction of oxygen atoms in the low temperature region tends to be more stereoscopic with trans- than with cis-olifins.'

This modification results in a reduction in length of the abstract by the removal of the four words. This reduction causes the improved abstract to represent .165 of the length instead of .167. This reduction in length changes the value of the data coefficient to 0.953. This modification does bring the abstract closer to the minimum level of acceptability, but there is still a need for additional improvements in this abstract.

The second example is the abstract of the article "Storage Organization in Programming Systems" which appeared in the Communications of the Association for Computing Machinery. The original abstract, shown in Figure 5.6, has a data coefficient of 1.094. This abstract can be further improved by application of the combination of sentences

ADDITION OF OXYGEN TO OLEFINS AT LOW TEMPERATURE IV REARRANGEMENTS. #R. KLEIN AND M.D. SCHEER, JOURNAL OF PHYSICAL CHEMISTRY 74(3), 613-616(1970). # A CONSIDERATION OF THE OXYGEN ATOM ADDITION TO CIS- AND TRANS-2-BUTENE IN THE TEMPERATURE REGION 77 TO 113 K LED TO THE FORMULATION OF A NEW TRANSITION INTERMEDIATE. IN THIS INTERMEDIATE, THE OXYGEN ATOM IS REPRESENTED AS BOUND IN A LOOSE, THREE-MEMBERED RING WITH, AND IN THE PLANE OF, THE OLEFINIC STRUCTURE OF THE REACTANT. OBSERVATIONS ON 2-BUTENES HAVE BEEN EXTENDED TO SEVERAL MORE STRAIGHT-CHAIN, INTERNAL OLEFINS IN THE LOW-TEMPERATURE REGION. COMPARISON OF THE TRANS-EPOXIDE TO KETONE RATIONS FROM THE CIS- VS THE TRANS-OLEFIN WITH INCREASING SIZE OF THE OLEFIN INDICATES THAT THESE RATIOS DIVERGE. REACTION OF OXYGEN ATOMS IN THE LOW-TEMPERATURE REGION TENDS TO BE MORE STEROSCOPIC WITH TRANS- THAN WITH CIS-OLEFINS. CARBONYL COMPOUNDS CONSTITUTE A SIZEABLE FRACTION OF THE PRODUCTS OF THE OXYGEN ATOM ADDITION TO OLEFINS IN THE LOW-TEMPERATURE REGION AND, AS HAS BEEN NOTED, AN INTRAMOLECULAR GROUP MIGRATION IS REQUIRES FOR CARBONYL FORMATION. THE PRINCIPAL CARBONYL PRODUCT IN THE TRANS-2-BUTENE REACTION AT 90 K IS 2-BUTANONE. THE FORMATION OF THIS KETONE REQUIRES THE MIGRATION OF H. COMPARED TO THE MIGRATION OF THE METHYL GROUP, THAT OF H IS SLIGHTLY FAVORED. CIS-2-BUTENE IS NOT USEFUL FOR THE COMPARISON, AS BOTH OF THE HYDROGEN ATOMS ATTTACHED TO THE OLEFINIC CARBON PAIR ARE SUPPRESSED THROUGH INTERACTION WITH OXYGEN IN THE COMPLEX. THE RELATIVE QUANTITIES OF 2-BUTANONE TO ISOBUTYRALDEHYDE IS TAKEN AS A MEASURE OF THE RATIO OF MIGRATION OF THE HYDROGEN ATOM TO THE METHYL GROUP. REACTIONS WERE EFFECTED AT 90 K IN THE APPARTATUS ROUTINELY USED FOR THIS PURPOSE. THE OLEFINS WERE DILUTED 10 TO 1 WITH PROPANE. THE EXPOSURE TIME OF OXYGEN ATOMS WAS 5 MINUTES, AND ABOUT 1% OF THE OLEFIN WAS REACTED. THE PRODUCTS WERE DETERMINED AT 135 AND A HELIUM FLOW OF 100 CC/MINUTE. THE CIS AND TRANS ISOMERS OF 3,4-EPOXY-3,4-DIMETHYLHEXANE WERE NOT SEPARABLE. LOCALIZTION OF THE OXYGEN ATOM IN THE TRANSITION COMPLEX PRECEDING ALKYL GROUP REARRANGEMENT IS NOT IN ACCORD WITH THE EXPERIMENTAL RESULTS. AT 90 K, THE RATIO OF ADDITION TO C-2 IS 16 TIMES THAT TO C-3. FOR MEP, ADDITION OF THE OXYGEN ATOM TO THAT CARBON ATOM OF THE DOUBLE BOND TO WHICH THE TWO METHYL GROUPS ARE ATTACHED WOULD BE EXPECTED TO BE FAVORED.

Document	Abstract
Total Words 2201	Total Words 364
Data Elements 146	Data Elements 23
	Content .1575
	Length .1672
	DC 0.953

Figure 5.5 Improved computer-produced abstract of "Addition of Oxygen Atoms to Olefins at Low Temperatures"

STORAGE ORGANIZATION IN PROGRAMMING SYSTEMS #JANE G. JODEIT: COMMUNICATIONS OF THE ACM. INTRODUCTION. IN THIS PAPER A REPRESENTATION OF DATA AND PROGRAMS IN STORAGE THAT CONTRIBUTES ORGANIZATIONAL SIMPLICITY, CODING CONVENIENCE, AND FUNCTIONAL VERSATILITY IN PROGRAMMING SYSTEMS IS DESCIBED. HERE PROGRAMMING SYSTEM MEANS THE REALIZATION OF A PROBLEM SOLUTION ON A COMPUTER, ANYTHING FROM MATHEMATICAL ANALYSIS TO LANGUAGE TRANSLATION. A PROBLEM SOLUTION IS DEFINED BY A COLLECTION OF ENTITIES, PROGRAMS AND DATA ITEMS SPECIFICALLY, THE GENERIC TERM FOR SUCH AN ENTITY IS AN ARRAY. EACH ARRAY IS NAMED AND CONTAINS AS ELEMENTS DATA OR SUBARRAYS. THE SET OF BRANCHES FROM A SINGLE SOURCE IS REPRESENTED BY A BLOCK CONTAINING CODEWORDS OR DATA AS APPROPRIATE. A CODEWORD WHICH CORRESPONDS TO A SIMPLE NAME, AS A ABOVE, BUT NOT IS CALLED A PRIMARY CODEWORD. ALL SUBARRAYS AND DATA ELEMENTS OF AN ARRAY ARE ADDRESSED "RELATIVE" TO THE SIMPLE NAME. THIS JUST MEANS THAT THE MTH ELEMENT OF THE NTH SUBARRAY IN THE ARRAY DATA IS NAMED DATA(N,M); IT HAS NO OTHER DESIGNATION. THE SET OF PRIMARY CODEWORDS THEN COMPLETELY CATALOGS THE ENTITIES OF A PROGRAMMING SYSTEM AND ALL ADDRESSING IS DONE THROUGH THESE CODEWORDS. THE OPERATING SYSTEM PROVIDES DYNAMIC ALLOCATION OF BLOCKS AND MAINTENANCE OF CODEWORDS. PRIMARY CODEWORDS NEVER MOVE , AND THE ADDRESSING IS INDEPENDENT OF SYSTEM COMPOSITION AND STORAGE ALLOCATION. CODEWORDS AS BLOCK LABELS AND THEIR USE IN ADDRESSING. A SET OF CONSECUTIVE STORAGE LOCATIONS IS CALLED A MEMORY BLOCK. EVERY SUCH BLOCK IS LABELED BY A SINGLE WORD CALLED A CODEWORD. AS REALIZED ON THE RICE UNIVERSITY COMPUTER THE GENERAL CODEWORD FORMAT IS SHOWN IN FIGURE 1, WHERE L IS THE LENGTH OF THE BLOCK LABELED BY THE CODEWORD C; I IS THE RELATIVE ADDRESS OF THE FIRST WORD IN THE BLOCK LABELED BY C; THE PORTION OF A CODEWORD USED IN INDIRECT ADDRESSING IS DESIGNED TO BE USED WITH THE HARDWARE DEFINITION OF THE RICE COMPUTER. IF *I IS ON, RETURN TO STEP FOR CODEWORD C+I+1 AT LEVEL I+1. IF *I IS NOT ON, USE CI+1 AS FINAL ADDRESS AND DO NOT ITERATE. THE LOCATION OF VT AND THE ORDER OF VT ENTRIES IS A FUNCTION OF SYSTEM COMPOSITION. INITIAL LOADING OF PROGRAMS AND DATA IS JUST A SEQUENCE OF ACTIVATIONS, AND THE BLOCKS WILL BE SEQUENTIALLY LOCATED IN THE STORAGE DOMAIN. AS A RUN PROGRESSES, BLOCKS MAY BE INACTIVATED AND NEW ONES ACTIVATED, SO THE GENERAL STATE OF THE STORAGE DOMAIN IS A MIXTURE OF ACTIVE AND INACTIVE BLOCKS. EACH ACTIVE BLOCK IN THE STORAGE DOMAIN IS LABELED BY A CODEWORD, WHICH MAY ITSELF BE A WORD IN AN ACTIVE BLOCK OF CODEWORD. ALREADY THE STORAGE CONTROL OPERATIONS OF BLOCK CREATION TO FORM ARRAYS AND FREEING OF ARRAYS HAVE BEEN MENTIONED. THE IMPLEMENTATION ON THE RICE COMPUTER PROVIDES A REPRESENTATION IN PRIMARY STORAGE WHICH IS IMMEDIATELY APPLICABLE THROUGH A HIERARCHY OF STORAGE DEVICES. THE DESCRIPTIVE PROPERTIES OF CODEWORDS, THE MODULARITY OF ARRAY STORAGE, AND THE PROTECTION POTENTIAL IN THE SYSTEM ALLOW THE CODEWORD STORAGE ORGANIZATION TO BE APPLIED IN A MULTIPROGRAMMING ENVIRONMENT. AN INTERRUPT WOULD ALLOW INTERVENTION FOR RETRIEVAL. STRUCTURED ARRAYS HAVE BEEN DESIGNED FOR SECONDARY STORAGE FILES.

Document	Abstract
Total Words 3680	Total Words 500
Data Elements 296	Data Elements 44
	Content .1486
	Length .1359
	DC 1.094

Figure 5.6 Computer-produced abstract of "Storage Organization in Programming Systems"

by means of a subordinate conjunction rule and the graphical reference rule. Application of these rules results in the abstract shown in Figure 5.7. In the original abstract the following two sentences appeared.

'The generic term for such an entity is an array.'

'Each array is named and contains as elements data or subarrays.'

These two sentences can be combined by application of the combination of sentences by means of a subordinate conjunction rule to generate the following sentence.

'The generic term for such an entity is an array, which is named and contains as elements data or subarrays.'

A similar application of the rule combines the two sentences

'A set of consecutive storage locations is called a memory block.'

'Every such block is labeled by a single word called a codeword.'

to form the following sentence:

'A set of consecutive storage locations is called a memory block, which is labeled by a single word called a codeword.'

The graphical reference rule should be applied to the following sentence.

'As realized on the Rice University computer the general codeword format is shown in Figure 1, ...'.

Since the figure is not included in the abstract the sentence would be modified to read

'As realized on the Rice University computer the general codeword format is shown in a figure, ...'.

STORAGE ORGANIZATION IN PROGRAMMING SYSTEMS #JANE G. JODEIT: COMMUNICATIONS OF THE ACM# IN THIS PAPER A REPRESENTATION OF DATA AND PROGRAMS IN STORAGE THAT CONTRIBUTES ORGANIZATIONAL SIMPLICITY, CODING CONVENIENCE, AND FUNCTIONAL VERSATILITY IN PROGRAMMING SYSTEMS IS DESCRIBED. HERE PROGRAMMING SYSTEM MEANS THE REALIZATION OF A PROBLEM SOLUTION ON A COMPUTER, ANYTHING FROM MATHEMATICAL ANALYSIS TO LANGUAGE TRANSLATION. A PROBLEM SOLUTION IS DEFINED BY A COLLECTION OF ENTITIES, PROGRAMS AND DATA ITEMS SPECIFICALLY. THE GENERIC TERM FOR SUCH AN ENTITY IS AN ARRAY, WHICH IS NAMED AND CONTAINS AS ELEMENTS DATA OR SUBARRAYS. THE SET OF BRANCHES FROM A SINGLE SOURCE IS REPRESENTED BY A BLOCK CONTAINING CODEWORDS OR DATA AS APPROPRIATE. A CODE WORD WHICH CORRESPONDS TO A SIMPLE NAME, AS A ABOVE, BUT NOT (A,I) IS CALLED A PRIMARY CODEWORD. ALL SUBARRAYS AND DATA ELEMENTS OF AN ARRAY ARE ADDRESSED "RELATIVE" TO THE SIMPLE NAME. THIS JUST MEANS THAT THE MTH ELEMENT OF THE NTH SUBARRAY IN THE ARRAY DATA IS NAMED DATA(N,M); IT HAS NO OTHER DESIGNATION. THE SET OF PRIMARY CODEWORDS THEN COMPLETELY CATALOGS THE ENTITIES OF A PROGRAMMING SYSTEM AND ALL ADDRESSING IS DONE THROUGH THESE CODEWORDS. THE OPERATING SYSTEM PROVIDES DYNAMIC ALLOCATION OF BLOCKS AND MAINTENANCE OF CODEWORDS. PRIMARY CODEWORDS NEVER MOVE, AND THE ADDRESSING IS INDEPENDENT OF SYSTEM COMPOSITION AND STORAGE ALLOCATION. A SET OF CONSECUTIVE STORAGE LOCATIONS IS CALLED A MEMORY BLOCK, WHICH IS LABELED BY A SINGLE WORD CALLED A CODEWORD. AS REALIZED ON THE RICE UNIVERSITY COMPUTER THE GENERAL CODEWORD FORMAT IS SHOWN IN A FIGURE, WHERE L IS THE LENGTH OF THE BLOCK LABELED BY THE CODEWORD C; I IS THE RELATIVE ADDRESS OF THE FIRST WORD IN THE BLOCK LABELED BY C. THE PORTION OF A CODEWORD USED IN INDIRECT ADDRESSING IS DESIGNATED TO BE USED WITH THE HARDWARE DEFINITION OF THE RICE COMPUTER. IF *I IS ON, RETURN TO STEP (1) FOR CODEWORD CI+1 AT LEVEL I+1. IF *I IS NOT ON, USE CI+1 AS FINAL ADDRESS AND DO NOT ITERATE. THE LOCATION OF VT AND THE ORDER OF VT ENTRIES IS A FUNCTION OF SYSTEM COMPOSITION. INITIAL LOADING OF PROGRAMS AND DATA IS JUST A SEQUENCE OF ACTIVIATIONS, AND THE BLOCKS WILL BE SEQUENTIALLY LOCATED IN THE STORAGE DOMAIN. AS A RUN PROGRESSES, BLOCKS MAY BE INACTIVATED AND NEW ONES ACTIVATED, SO THE GENERAL STATE OF THE STORAGE DOMAIN IS A MIXTURE OF ACTIVE AND INACTIVE BLOCKS. EACH ACTIVE BLOCK IN THE STORAGE DOMAIN IS LABELED BY A CODEWORD, WHICH MAY ITSELF BE A WORD IN AN ACTIVE BLOCK OF CODEWORD. ALREADY THE STORAGE CONTROL OPERATIONS OF BLOCK CREATION TO FORM ARRAYS AND FREEING OF ARRAYS HAVE BEEN MENTIONED. THE IMPLEMENTATION OF THE RICE COMPUTER PROVIDES A REPRESENTATION IN PRIMARY STORAGE WHICH IS IMMEDIATELY APPLICABLE THROUGH A HIERARCHY OF STORAGE DEVICES. THE DESCRIPTIVE PROPERTIES OF CODEWORDS, THE MODULARITY OF ARRAY STORAGE, AND THE PROTECTION POTENTIAL IN THE SYSTEM ALLOW THE CODEWORD STORAGE ORGANIZATION TO BE APPLIED IN A MULTIPROGRAMMING ENVIRONMENT. AN INTERRUPT WOULD ALLOW INTERVENTION FOR RETRIEVAL. STRUCTURED ARRAYS HAVE BEEN DESIGNED FOR SECONDARY STORAGE FILES.

Document	Abstract
Total Words 3680	Total Words 497
Data Elements 296	Data Elements 44
	Content .1486
	Length ..1351
	DC 1.106

Figure 5.7 Improved computer-produced abstract of "Storage Organization in Programming Systems"

These improvements result in a reduction in length from .136 of the original article to .135 and an improvement of the data coefficient from 1.094 to 1.106. The abstract also seems to be more cohesive and unified without as many short, disjoint sentences.

The third example illustrates that in some cases the application of the improvement procedures will not increase the value of the data coefficient, but the value of the data coefficient will not decrease. The third sample abstract, shown in Figure 5.8, is of the article, "Mini-computers Turn Classic" which appeared in Data Processing. Only one rule, the combination of sentences with identical headwords of the subject noun phrases and identical main verb phrases by means of a coordinate conjunction, was applicable. This rule was applied to the last two sentences of the abstract which are

> 'Individual manufacturers offer ALGOL, BASIC, and FOCAL compilers.'
>
> 'Cost manufacturers offer programming support on an individually negotiated contract basis.'

to yield the sentence

> 'Individual and cost manufacturers offer ALGOL, BASIC, and FOCAL compilers, and programming support on an individually negotiated contract basis, respectively.'

The resulting sentence is not any shorter than the two original sentences, hence there is no reduction in length. There is no reduction in content so the data coefficient remains the same. The modified abstract is shown in Figure 5.9.

These examples serve to illustrate that the application of procedures to modify the sentences of the abstract by structural analysis

MINICOMPUTERS TURN CLASSIC. * J.J.BARTIK:DATA PROCESSING.12(1), 42-50(1970).
MANUFACTURERS EVEN OFFER CENTRAL PROCESSORS WITH NO MEMORY WHATSOEVER. DATA IS TRANSFERRED BETWEEN MEMORY AND THE CENTRAL PROCESSOR VIA A MEMORY BUS. THE ENTIRE CORE IS ADDRESSABLE VIA INDEXING OR INDIREST ADDRESSING AND GENERALLY INPUT/ OUTPUT MINICOMPUTERS INCLUDE A PROGRAMED PARTY LINE I/O CHANNEL. THE DATA CHANNEL IS ONE WORD WIDE, EIGHT BITS FOR AN 8-BIT PROCESSOR AND 16 BITS FOR A 16-BIT PROCESSOR. SLOW SPEED, CHARACTER-ORIENTED DEVICES ALSO INTERFACE TO THE CHANNEL FOR DATA TRANSFERS, AND EACH TRANSFER IS UNDER PROGRAM CONTROL. THE BLOCK TRANSFER OPTION DOES ALLOW RELATIVELY HIGH-SPEED DEVICES TO INTERFACE TO THE CHANNEL FOR DATA TRANSFERS. A NUMBER OF MANUFACTURERS ALSO OFFER A GENERAL-PURPOSE INTERFACE TO THE PROGRAMED I/O CHANNEL TO PROVIDE FOR ADDRESSING AND CONTROLLING SPECIAL-PURPOSE I/O DEVICES AND FOR TRANSFERRING DATA.
MOST MINICOMPUTERS CAN SUPPORT AN OPTIONAL DIRECT MEMORY ACCESS CHANNEL.
INDIVIDUAL MANUFACTURERS OFFER ALGOL, BASIC, AND FOCAL COMPILERS.
COST MANUFACTURERS OFFER PROGRAMMING SUPPORT ON AN INDIVIDUALLY NEGOTIATED BASIS.

Document	Abstract
Total Words 2430	Total Words 162
Data Elements 194	Data Elements 12
	Content .0619
	Length .0667
	DC 0.928

Figure 5.8 Computer-produced abstract of "Mini-computers Turn Classic".

MINICOMPUTERS TURN CLASSIC. # J.J.BARTIK:DATA PROCESSING.12(1), 42-50(1970).
TWOMANUFACTURERS EVEN OFFER CENTRAL PROCESSORS WITH NO MEMORY WHATSOEVER. DATA IS TRANSFERRED BETWEEN MEMORY AND THE CENTRAL PROCESSOR VIA A MEMORY BUS. THE ENTIRE CORE IS ADDRESSABLE VIA INDEXING OR INDIREST ADDRESSING AND GENERALLY INPUT/ OUTPUT MINICOMPUTERS INCLUDE A PROGRAMED PARTY LINE I/O CHANNEL. THE DATA CHANNEL IS ONE WORD WIDE, EIGHT BITS FOR AN 8-BIT PROCESSOR AND 16 BITS FOR A 16-BIT PROCESSOR. SLOW SPEED, CHARACTER-ORIENTED DEVICES ALSO INTERFACE TO THE CHANNEL FOR DATA TRANSFERS, AND EACH TRANSFER IS UNDER PROGRAM CONTROL. THE BLOCK TRANSFER OPTION DOES ALLOW RELATIVELY HIGH-SPEED DEVICES TO INTERFACE TO THE CHANNEL FOR DATA TRANSFERS. A NUMBER OF MANUFACTURERS ALSO OFFER A GENERAL-PURPOSE INTERFACE TO THE PROGRAMED I/O CHANNEL TO PROVIDE FOR ADDRESSING AND CONTROLLING SPECIAL-PURPOSE I/O DEVICES AND FOR TRANSFERRING DATA.
MOST MINICOMPUTERS CAN SUPPORT AN OPTIONAL DIRECT MEMORY ACCESS CHANNEL.
INDIVIDUAL AND COST MANUFACTURERS OFFER ALGOL, BASIC, AND FOCAL COMPILERS AND PROGRAMMING SUPPORT ON AN INDIVIDUALLY NEGOTIATED CANTRACT BASIS, RESPECTIVELY.

Document	Abstract
Total Words 2430	Total Words 162
Data Elements 194	Data Elements 12
	Content .0619
	Length .0667
	DC 0.928

Figure 5.9 Improved computer-produced abstract of "Mini-computers Turn Classic".

can improve the quality of the abstracts and this improvement will result in the same or increased value of the data coefficient. These procedures can be applied to all abstracts and although the improvement may be more noticeable in some abstracts than others, the quality will not be lowered in any example. This set of rules is presently not sufficient to improve abstracts so that the data coefficient is above 1.0 for all examples. Therefore additional rules should be added to the modification procedure.

2. Improvement of the System Implementation

The abstracting system, ADAM, is, at present, capable of producing abstracts from the input of a complete document. ADAM can be improved by improving the quality of the abstracts (see Section 1) and by improving the implementation of the abstracting algorithms. The manner in which the algorithms are programmed and the efficiency of operation of the programs will determine the feasibility of actual use of ADAM in an operational environment. The system improvements discussed in Section 2 are aimed at efforts to make ADAM operate as efficiently as possible and to make it competitive with existing abstracting procedures.

2.1. Modification of the Word Control List

As has already been pointed out, ADAM consists of two basic components, the abstracting rules and the Word Control List (WCL). The WCL is a list of words and word strings with which are associated codes indicating the semantic and/or syntactic role each plays. The dicotomy of rule and WCL is an important system design parameter. By making the

WCL act as an interface between the text and the abstracting rules, processing complexity is reduced, efficiency is improved and considerable flexibility is gained in the control one has over the way in which the abstracting system operates. The abstracting rules deal only with metalinguistic (nonterminal) symbols. The WCL supplies these symbols. Consequently, if we can assume that the rules are adequate, the WCL will determine the exact nature and content of an abstract produced by the system. It is therefore very important to know as precisely as possible what entries should be in the WCL, what codes should be associated with each entry, how frequently each entry is actually applied in the processing of an "average" document and so on. In short, it is necessary, both for effective and efficient abstract production, to determine goodness criteria for the WCL and to use these criteria as a basis for optimizing the WCL.

Recalling that the design of ADAM is based on the notion that expressions signifying low information content are more nearly constant and are therefore more easily identified, it is readily understood why it is important to construct the WCL so that it incorporates as high a concentration of these well-used expressions as possible. To develop such a WCL, several aspects of the existing WCL should be studied. These aspects include the determination of the frequency of use of each entry in the existing WCL, an analysis of abstracts derived from a set of documents to determine additional WCL entries which might profitably be used, and a careful determination of the proper semantic and syntactic roles to be associated wtih each WCL entry. These studies

should also be designed to answer the question of whether additional codes should be associated with WCL entries. Finally, a procedure should be implemented through which the WCL may be modified as dictated by any particular application of the abstracting system.

2.2. Use of the Word Control List to Control the Subject Orientation of Abstracts

As initially designed, ADAM was viewed as being able to produce abstracts with no special regard for the subject area of the original document. While I still hold this view, it must be recognized that it is clearly possible to produce abstracts slanted toward some particular subject area, for some particular purpose. In ADAM this control over the subject orientation of the produced abstract could be gained through manipulation of the WCL. And such manipulation can be done without any modification of the programs of the system.

Modifications of the WCL would fall into one or the other of two classes: 1) addition of words common to a particular subject area, which thus serve as function words, and 2) addition of special entries with semantic code I (see Table 3.1), which would cause sentences containing these entries always to be selected for the abstract. Such modifications would have the effect of providing to the reader (or searcher) more specific data (i.e., would produce informative rather than indicative abstracts) and of indicating the viewpoint of the abstracting system.

2.3. Analysis of Data Structures

A data structure may be defined as an ordered set of data elements, together with some particular interpretation. The interpretation assigns to a particular data element in a particular position in the structure some particular function or meaning. A data structure may also be characterized in terms of the procedure which utilizes it. In the case of ADAM, several data structures are employed for the several basic procedures which make up the system. Three important data structures utilized in the abstracting system may be identified. These are

1. the structure associated with the input text,
2. the data structure associated with the WCL and the allied matching process, and
3. the data structure utilized in the application of the abstracting rules.

The existing data structures should be analyzed in terms of their overall efficiency in the abstracting process. Alternative data structures which may prove more efficient than those presently in use, should be considered.

2.3.1. The Structure Associated with the Input Text

In the case of the data structure associated with the input of the original document, ADAM utilizes a memory area of fixed size for storing the text during the entire process of producing an abstract from it. The storage allocated for this purpose (40,000

characters) is adequate for the storage of documents which contain, on the average, fewer than 5,000 text tokens. If the length of the document exceeds this number of tokens, ADAM prints an error message, skips over the document and thus does not produce an abstract for that document. It is clear that the existing data structure is therefore inefficient for several reasons. It is obviously inefficient in dealing with large documents, since such documents would have to be reduced in size and recycled through the system. The data structure is also inefficient because storage is wasted for all documents smaller than the maximum document size allowed. Since documents are of variable length, a better data structure might be one which provides just the amount of storage needed to store a document (provided the total amount of storage available in the computer system is not exceeded). But such a data structure requires a dynamic storage allocation mechanism, and the time required to manage the storage allocation might offset any gain in efficiency of use of storage. Various methods for dynamically allocating storage for the input document should be studied along with a comparison of the overall efficiency of these methods with the static storage allocation method currently employed in the system.

To exemplify the allocation of storage areas for document input, consider the following. The average document which the abstracting system will come into contact with would have a length equivalent to 15,000 characters. On the basis of this estimate, the system would be designed so that this number of characters was allocated before document input was initiated. If, during the actual input step, this

initial amount of storage was found to be insufficient, an additional block of 1,000 characters of storage would be allocated. If this additional space was not sufficient to hold the entire document, an additional block would be allocated. This process would be continued until the document had been read in its entirety, or until no more storage was available for allocation. Such a technique is not, however, without its problems. Since each new block of storage allocated will in all probability not be continguous with the previously allocated block, a storage management routine would have to be provided to keep account of the storage addresses associated with each block and to provide a chaining mechanism for handling the blocks as an integral unit. Furthermore, 1000-character blocks might still result in somewhat inefficient usage of the memory, but smaller blocks would increase the storage management problems. Thus it should be clear that optimizing the data structure for document storage requires careful study of alternative methods for storage allocation and management.

2.3.2. The Structure Associated with the Word Control List and the Matching Process

In the prototype system, the Word Control List is stored on disk in alphabetical order and is read into core memory at the time it is to be compared with the text of the original document. Since the WCL is in alphabetical order, the text must also be alphabetized. This is accomplished by means of a pointer sort using the method of quadratic selection with exchange. Obviously if a way could be found of effecting the comparison of WCL and text without having to sort the text then a

considerable saving in processing time would be gained.

Another problem associated with data representation in the WCL is that each entry must match exactly a text word or word string or else the entry is considered to be a complete mismatch. But many words in the text may have the same stem (or root) as an entry in the WCL, but they are encumbered with inflectional endings which cause them to appear different from the WCL entry. To cope with this difficulty the WCL is now made to contain the various forms of the same word (as, for example, UNUSUALLY and UNUSUAL). From other studies (9) it is known that it is possible in most instances to deal with inflectional forms by means of affix elimination techniques. But more recently, it has been shown that key generation techniques usually associated with hashing studies may provide an ever more attractive solution to the word form problem (10). Furthermore, such a method suggests a solution to the matching problem alluded to earlier. A text word could, by suitable means, be converted to a key (which would be used in the actual matching operation) which would be converted to an address pointing to that entry or set of entries in the WCL with which the word might match. Matching could then be carried out on the key only, or first on the key and then the entire word. Certainly, such data-controlled addressing has been used successfully in systems dealing with much larger data structures than that associated with the WCL (11).

2.3.3. The Structure Utilized in the Application of the Abstracting Rules

The third data structure of major importance in ADAM is that called the Table. The Table consists of entries corresponding to the words in the input document. Each Table entry occupies 8 characters of storage and contains the address of each word in the input document, the length of the word, an alphabetic pointer, and space for a semantic and a syntactic code. A fixed amount of storage is allocated for the Table in advance of its creation, and every word in the input text has a corresponding Table entry. The reason for this is that the input text is compared with the WCL through use of the pointers contained in the Table. While the Table is an essential component of the abstracting system, it can be made more efficient, at least in terms of storage utilization. If a way can be found of effecting matching with the WCL without the necessity of alphabetizing the text, then the Table could be built to contain only those text words which matched WCL entries, together with sentence and clause markers (periods and commas, mainly), without detriment to the operation of the abstracting procedures. One might also gain some additional reduction in storage requirements by representing the semantic and syntactic codes numerically rather than by character codes.

2.4. Design of the Semantic Module

ADAM is programmed with several modules to perform specified functions which are controlled by one main module. The abstracting rules which implement the functions specified in the Word Control List

are programmed in the Semantic module. This module can be reprogrammed to incorporate new abstracting rules without changing the rest of the system. Although most minor modifications can be made by changes in the Word Control List, I feel that two major changes should be incorporated into the Semantic module.

First, the verb check within the Semantic module should be eliminated in favor of a verb check by means of structural analysis. The Semantic module currently has a section of code which examines the syntactic codes of each of the words in each sentence in order to locate a verb code. If the sentence contains at least one verb, then it is included in the abstract if it has met all other requirements for inclusion. If the sentence does not contain a verb then it is deleted. This procedure guarantees that each sentence in the abstract will contain a verb. While it is clearly desirable that each sentence of the abstract contain a verb, this method poses two disadvantages. First, the size of the Word Control List must be increased because of the addition of all the verbs that are to be recognized. Since processing time increases with increases in the size of the dictionary, this cuts down on the system throughput. Second, many valid verbs are unrecognized because they do not match an entry in the Word Control List. These sentences, which might be valuable in the abstracts, are thus omitted. Since most sentences which are selected for the abstract will be taken directly from the document, it is probably safe to assume that those sentences are well-formed and contain a verb. Thus, no verb check is needed in those cases. If a sentence is not well-formed, a simple

check for a set of specified verbs will not always serve to eliminate the sentence. A better approach would be to incorporate in the sentence modification phase of the program (which is discussed in Section 1 of this Chapter) a test for well-formed sentences which all include verbs. This modification would allow greater numbers of verbs to be recognized as valid and would increase the efficiency of the system.

Second, a semantic code for 'super-delete' should be added entirely within the Semantic module. This code would take precedence over the I code (see Table 3.1), which indicated the highest priority of importance. The super-delete code would be assigned to a sentence to indicate that it should not be reinstated by any sentence which refers to it. A super-delete code would be assigned to all questions, equations, or direct quotations. A super-delete code would also be used in the reinstatement procedure for intersentence reference. If a sentence requires an antecedent, then the previous sentence would normally be reinstated. If the previous sentence has been assigned a super-delete code, it would be impossible to reinstate the required antecedent sentence so then the sentence under examination would be marked with a super-delete code. This would be useful in preventing the inclusion of a sentence without an antecedent in the abstract and in preventing the reinstatement of that sentence if the following sentence also requires an antecedent. These two changes could be incorporated in the Semantic module and would improve both the performance and the efficiency of the abstracting system.

3. Directions for Future Research

The research described in this dissertation could serve as the basis for additional research projects which are beyond the scope of this project. There are five areas for further research: large-scale testing, linguistic analysis, system implementation, adaptive and learning behavior, and inter-system compatability. These five areas are discussed in the next five sections.

3.1. Large-Scale Testing

The automatic abstracting system and evaluation criterion appear to be ready for a test of their practicability in an operational environment. The test should be constructed as a realistic procedure for producing abstracts for an information system that currently produces abstracts manually. The goals of the existing system should be carefully specified so that the data element can be defined in accordance with the use of the abstracts produced. The cost of abstracting a given set of test documents should also be determined. Cost and quality comparisons should be made based on a given set of documents which are representative of the total data base.

The abstracting system should be implemented to produce abstracts as similar to the desired system goals as possible. This might mean modification, addition, or deletion of entries in the Word Control List. The format of the output should be changed to conform as closely as possible to the format of the output used by the system. The evaluation criteria should be programmed to recogn e the defined data

element and make the data content/ length comparisons. The test set of manually-produced abstracts should be available in machine-readable form for their evaluation by the same criterion.

When all the parameters are carefully specified, the next step is to actually perform the test. As with many experiments, the time needed to perform the test will be small compared to the time of preparation and evaluation of results. Careful records of time, memory requirements and cost should be made on the computer runs needed to abstract the documents and to perform the evaluation. The quality of the computer-produced abstracts should be compared to the quality of the manually-produced abstracts and to some absolute standard in terms of the data concentration factor. The results of this evaluation should be used to improve the efficiency and operation of the abstracting system in areas of deficiency. The cost and quality factors could be used to study the feasibility of system implementation.

The abstracting system may not give a high performance on its first large-scale test. Bernier states that beginning human indexers may not do too well on their first experience, either.

> Beginning indexers at Chemical Abstracts often had 50-75% of their index entries changed upon checking. Perhaps half of these changes were of a minor nature (e.g., for paraphrasing, abbreviation, elimination of redundancy) and not of a major nature (e.g., omissions, scattering, mistakes or errors). Had these indexers not received their B.S. or higher degrees in chemistry, then the percentage of changes would have been greater. (12)

A beginning indexer must be trained by an evaluation of his work with suggestions for improvement. It is only reasonable to expect that the

abstracting system will have to be modified to reflect its experience.

3.2. Linguistic Analysis

ADAM, as well as other natural language processing programs can be improved with an increased awareness of the fundamental nature of language. Most of the rules of the abstracting system are based on an *ad hoc* determination of what seemed reasonable. These rules could be substantiated or denied by studies of the way language is used to communicate ideas. For a given data base, statistical studies could be made on the frequency of appearance in the text of entries in the World Control List. Statistical studies could also be made on the number of times each rule in the semantic module is applied. The results of these studies could be used to eliminate the rules that are infrequently applied and that decrease the efficiency of the abstracting system. The results could also be used to indicate which rules process the most frequently used expressions.

The abstracting system could be further improved by incorporating additional algorithms which indicate semantic and syntactic information. Input of both upper and lower case characters would allow an analysis of capitalization in each sentence. This would be useful in identifying proper names and sentence boundaries. Incorporation of analysis programs to identify phrase and clause boundaries might also be useful in selecting portions of a sentence for inclusion in the abstract. This, along with more sophisticated algorithms to determine the use of commas and periods, would enable the program to have more complete

data on the manner in which the ideas in the document are expressed, both in words and in structure. It would be helpful to be able to recognize synonyms in the text. This synonymity might be between singular and plural forms of the same word or between words which are completely different in spelling. For example, it would be beneficial if the system could make use of the fact that "automatic abstracts", "automatically-produced abstracts", and "computer-produced abstracts" all refer to one idea.

In many documents, one complete thought is written in more than one sentence. Identification of intersentence references is very important in preserving the meaning and the coherency of the abstracts. Determination of intersentence references is dependent on both semantic and syntactic information from the document. With this information, it might be possible to create sentences for the abstract which combine two sentences from the document into a single sentence in the abstract. The sentences of the abstract should also then come together to create a coherent abstract.

3.3. System Implementation

The future of automatic abstracting as a viable method of producing abstracts in an operational environment depends in large measure on the reliability of the system. The abstracts produced should have a data coefficient above the minimum level of acceptability in almost every case. A high degree of dependability is necessary before any manager would even consider converting from a manual to a computer-

based method for production of abstracts..

Perhaps a possible method for implementing a computer-based system would be to include human post-editing of all abstracts produced. This would remove obvious errors and allow for maintenance of a quality standard. The initial systems that convert from manually-produced to computer-produced abstracts might have to retain all of their abstractors initially to edit the computer output. If the editors' remarks could be used as feedback to the system, the quality of the computer-produced abstracts could be improved.

3.4. Adaptive and Learning Behavior

Ideally ADAM should be able to learn from his mistakes and not be a novice abstractor forever. The ability to improve will probably result from the modification of the abstracting system components by the system designers. It would be desirable if ADAM were able to learn from his past experience and adapt his behavior based on the data he currently processes. This adaptive behavior might be based on a link with previous experience through a memory bank. It might also include access to other information systems which provide document services to users. This abstracting system, with its ability to learn, adapt, communicate, and remember could be truly described as an intelligent system.

This level of intelligence, and it is questionable if it is even desirable, certainly will be difficult, if not impossible to attain. This is an example of what Bar-Hillel calls the 80% fallacy. "The

remaining 20% will require not one quarter of the effort spent for the first 80%, but many, many times this effort, with a few percent remaining beyond the reach of every conceivable effort." (13) My research has been where the slope of the curve on the asymptotic approach to perfection has been relatively small. Further research will require a much greater expenditure of effort for even modest gains.

3.5. Inter-System Compatability

As has been pointed out elsewhere (14, 15, 16), an Information Storage and Retrieval system acts as an interface between a source (a scientist originating a document) and a receiver (a scientist seeking a document) as depicted in Figure 5.10. The operation of an Information Storage and Retrieval system involves a number of processes which are effected in approximately chronological order. Let us consider these main processes in this order. First, although an automatic abstracting system requires an entire document in computer-processable form, this fact should produce no limitation on the operation of an Information Storage and Retrieval system, since more and more primary sources are being published using computer-based photo-composition (computer output on microfilm, COM) with the full text as a by-product data base (17). An Information Storage and Retrieval system should be able to take full advantage of such data bases. Furthermore, optical character readers (OCR) are becoming technically more sound, operationally more general and reliable, and economically more attractive (18), so that original documents not in machine-readable form can be obtained in that

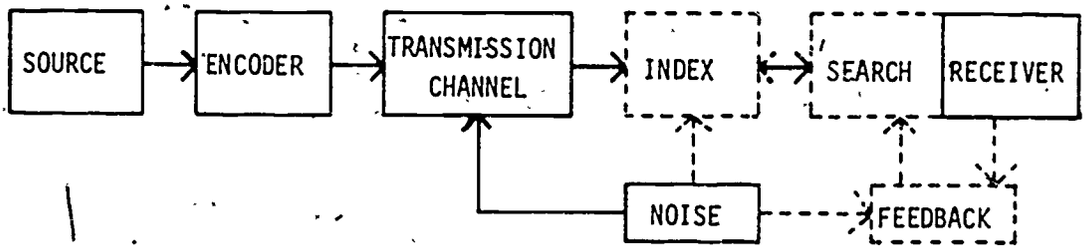


Figure 5.10 General communication system model showing the role of the Information Storage and Retrieval System as an interface between source and receiver. Components of the Information Storage and Retrieval system are identified by broken lines.

form rather easily.

The next step in the process is the automated editing of the input text. In general, this editing involves the deletion of unwanted portions of the text (for instance, figures or other document attributes considered to be unsuitable in a given application), or the reorganization or restructuring of some document attributes (such as standardization of bibliographic citations). Once the editing process is completed the document is ready to be abstracted. Simultaneously with abstract preparation, the references in the original document could be isolated as input to a citation indexing procedure. Citation indexes have considerable value in themselves as search tools (19, 20).

Once an abstract is available, it may be printed as part of an abstract bulletin and at the same time be used as the basis for an automated indexing procedure. A variety of indexing methods is already available (21, 22, 23, 24) which could be employed to produce a printed index from the abstract. A title index might also be produced as a smaller, less expensive, "throwaway" index. The index entries derived from the abstract could be used without further processing (free-vocabulary indexing) or the index vocabulary could be controlled automatically (as is done presently in at least one instance (25)).

But having created an index to the document collection, using the abstract as a basis, does not obviate the use of the abstract as the basis for an automated search and retrieval system, either batch or on-line (26). Furthermore, the abstract could be augmented with index terms derived through use of a vocabulary-control procedure. And a

variety of formats for the data suggest themselves, which could enhance the effectiveness of a retrieval system under particular circumstances.

Responses to a user inquiry to a retrieval system based on abstracts could take the form of the bibliographic citation, but a more "intelligent" system, possessing such a data base, could also provide parts of or whole abstracts as responses and, under suitable circumstances, parts or all of the original document could be supplied as responses as well.

The central feature of the hypothetical Information Storage and Retrieval system (summarized in Figure 5.11) that I have described is the abstract. With an efficient means of producing effective abstracts, the entire Information Storage and Retrieval system becomes feasible. Automated abstracting methods constitute the sole missing component of such a system. It is, therefore, one of the purposes of research and development work on automatic abstracting to provide this missing component and thereby make such an Information Storage and Retrieval system possible.

4. Conclusions

4.1. The Design of an Automatic Abstracting System

The manual production of abstracts is based on the ability of a trained abstractor to read a document, understand its contents, select certain key ideas, and rewrite these key ideas to form a coherent abstract. Figure 5.12 illustrates these basic steps in the manual production of abstracts. Automatic abstracting systems are designed

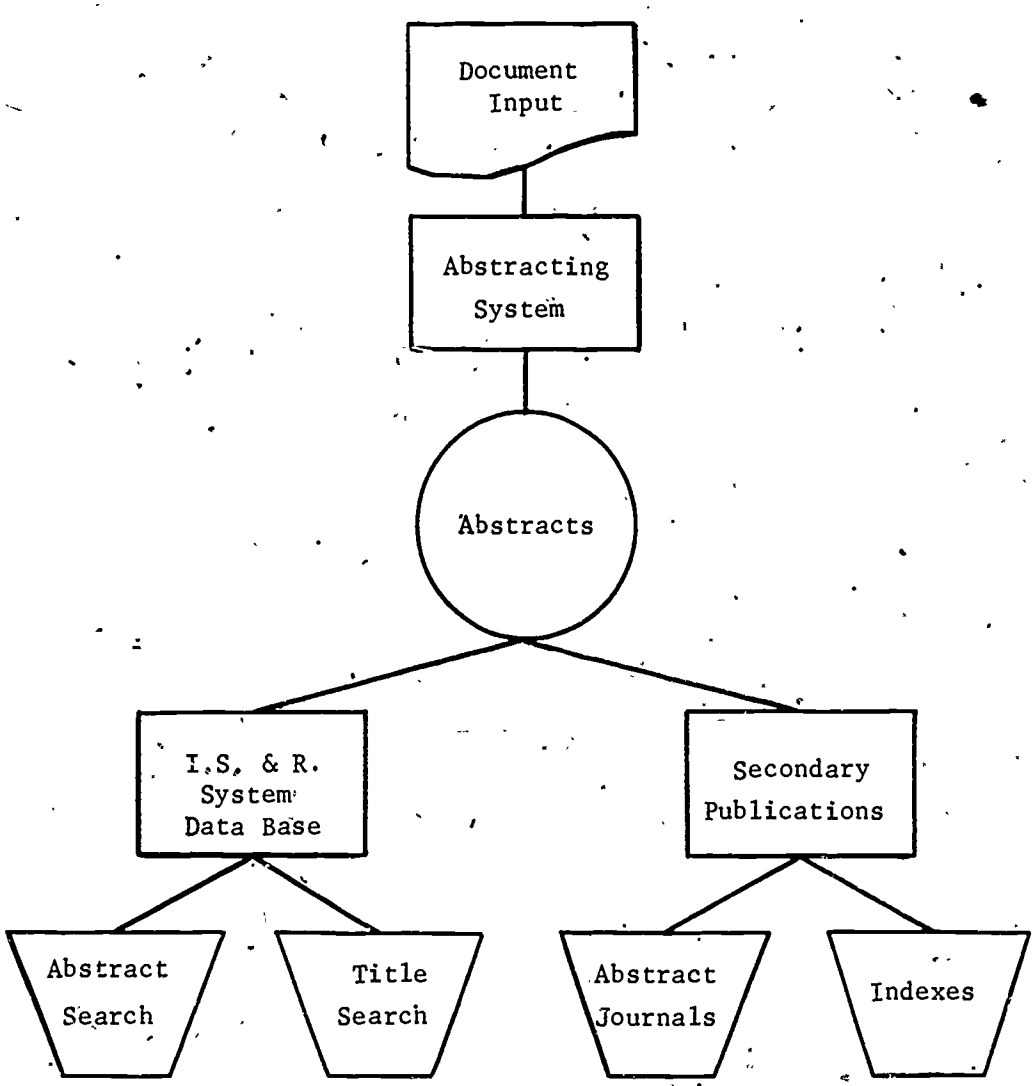


Figure 5.11 Hypothetical Information Storage and Retrieval system using abstracts as the basic data source.

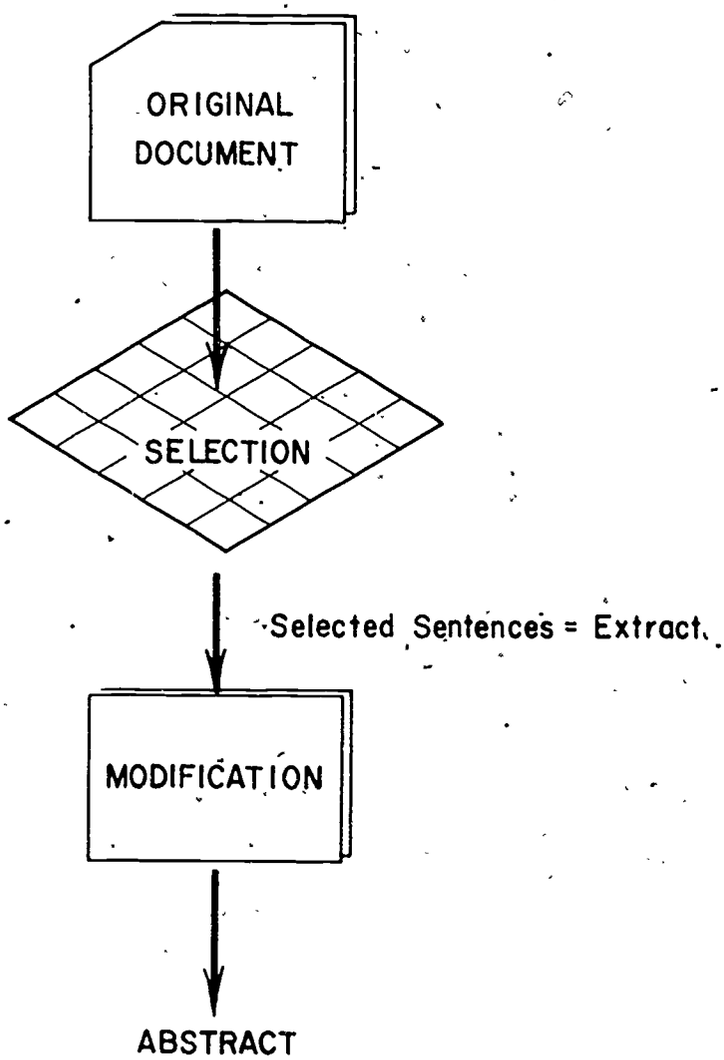


Figure 5.12 The basic steps in the manual production of abstracts.

to produce abstracts in a similar manner. The system receives an original document as input. The system then selects certain key sentences from the document to be included in the extract. Figure 5.13 shows the general character of all existing automatic abstracting systems. The selection method of the abstracting system described in this dissertation relies on the rejection of sentences which are unsuitable for inclusion in the extract. Those sentences which are not rejected are included in the final abstract. This method coincides with the intuitive idea that an abstract should help the user by screening out those portions of a document which are not the most useful to him. This method also provides a practical means of implementing the process of abstracting.

4.2. The Improvement of an Automatic Abstracting System

Comparison of Figures 5.12 and 5.13 suggests that an important refinement of automatic abstracting systems would be that involving the development of procedures for the modification of the form, arrangement, and content of the sentences selected for the abstract. This modification would produce abstracts in which the flow of ideas was improved and which represented a more nearly coherent whole. This modification is based on a structural analysis and revision of each sentence in order to make the abstracts more acceptable to the reader. The addition of the modification procedure is shown in Figure 5.14. The generation and revision of sentences is as complex a procedure as the selection procedure, so addition of this phrase will make the abstracting system

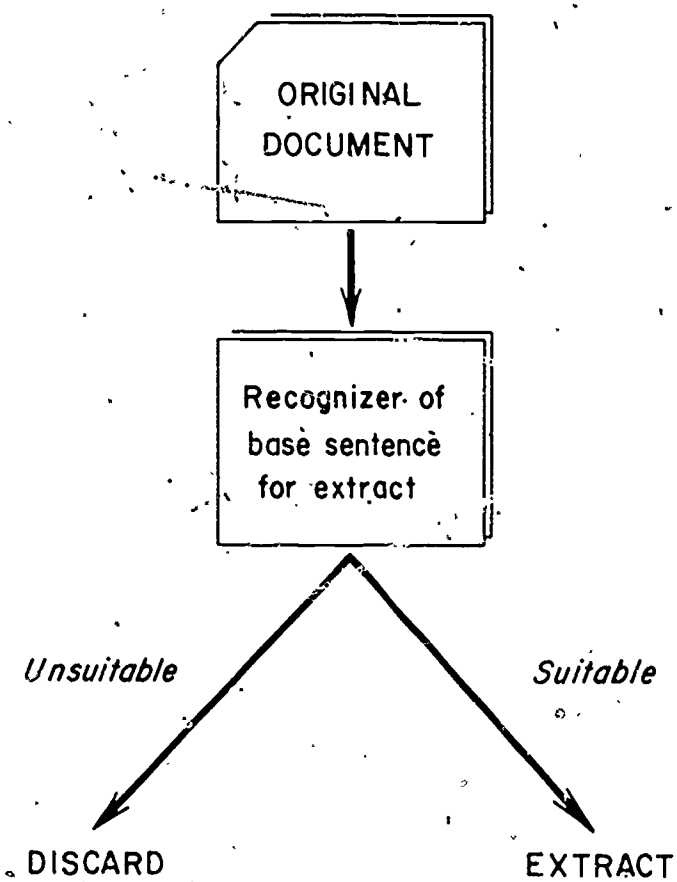


Figure 5.13 The basic steps in the computer production of abstracts.

larger and more costly to operate. The revision of sentences appears to be warranted because the resulting abstracts have a more readable, coherent style.

4.3. The Evaluation of an Automatic Abstracting System

The design of an automatic abstracting system must rely on some intuitive idea of what constitutes a good abstract. This criterion of quality must be explicitly defined so that it may be used as a measure of the effectiveness of the abstracting system. Inclusion of an evaluation procedure in the abstracting system is shown in Figure 5.15. The evaluation criterion that I have developed is based on the axiom that the best abstract among a set of abstracts is that one which presents a maximum amount of data in the minimum amount of length. Since measuring the "information content" of a document presents a very difficult practical problem, a measure of data, which is easier to implement, is formulated. Since information can be considered as data of value in decision making, the amount of information in a given abstract will always be less than or equal to the amount of data. This method of evaluating abstracts can be adapted to many different systems where abstracts are used by adapting the definition of data element to the goals of the particular system. This method also provides an objective criterion for abstract evaluation. Almost all previous evaluation techniques relied on the opinions of people which would often not be uniform or consistent. Algorithmic improvement of automatic abstracting systems using only a subjective method of

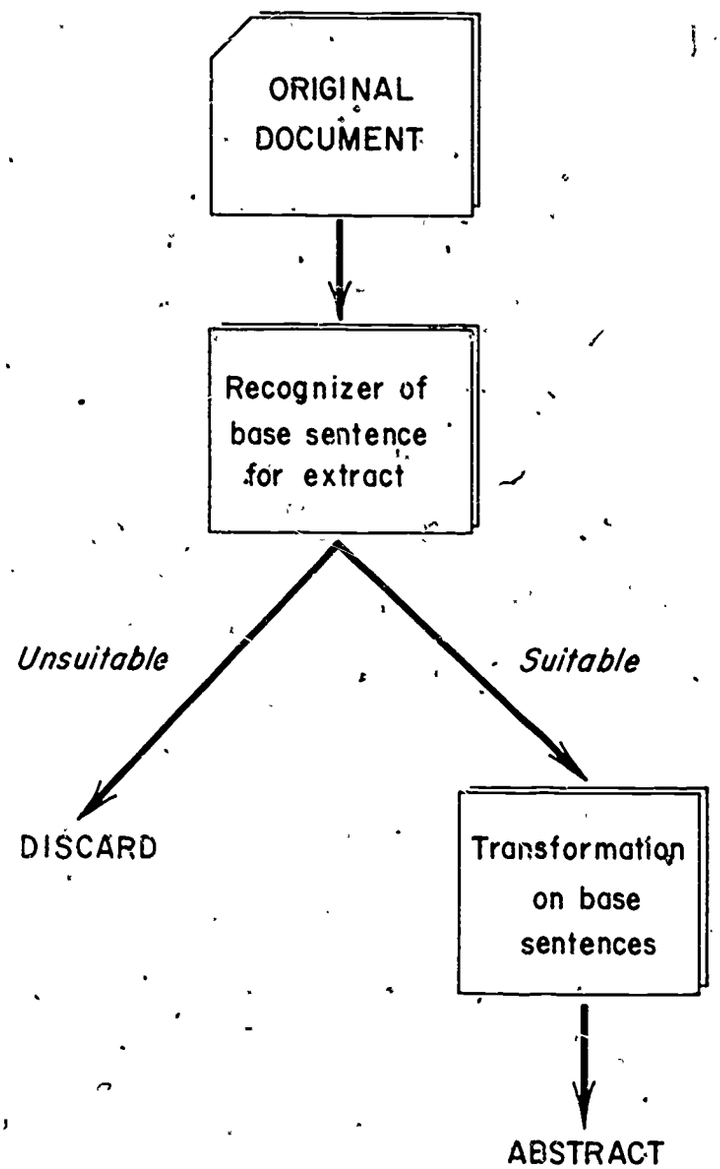


Figure 5.14 The addition of a modification procedure to the computer production of abstracts.

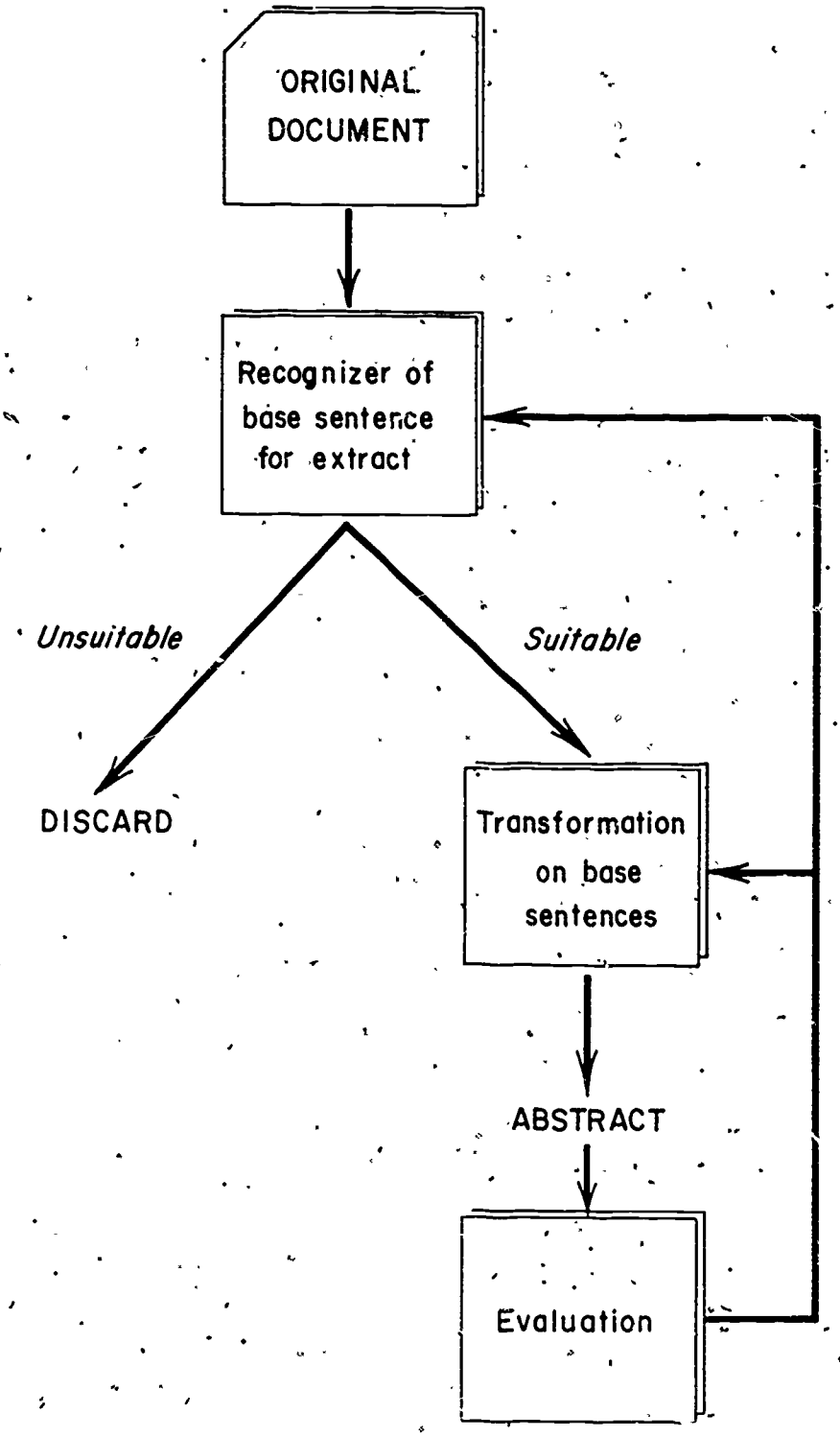


Figure 5.15 The addition of an evaluation procedure to the computer production of abstracts.

evaluation is almost impossible, so this development of an objective method is particularly important. This dissertation has presented my research on an automatic abstracting system, its design, evaluation, and improvement.

References

1. R. E. Wyllys, "Extracting and Abstracting by Computer," in H. Borko (ed.), Automated Language Processing, John Wiley and Sons, Inc., New York, New York, 1968, p. 128.

2. M. H. Needleman, Handbook for Practical Composition, McGraw-Hill Book Company, New York, New York, 1968.

3. C. C. Fries, The Structure of English, Harcourt, Brace & World, Inc., New York, New York, 1952.

4. H. Kucera and W. Francis, Computational Analysis of Present-Day American English, Brown University Press, Providence, Rhode Island, 1967.

5. S. Marvin, J. E. Rush and C. E. Young, "Grammatical Class Assignment Based on Function Words," in A. Berton (ed.), The Social Impact of Information Retrieval, Medical Documentation Service, Philadelphia, Pennsylvania, 1970.

6. W. A. Cook, S. J., "Case Grammar: From Roles to Rules," Languages and Linguistics Working Papers Number 1. Georgetown University Press, Washington, D.C., 1970.

7. C. E. Young, Design and Implementation of Language Analysis Procedures with Application to Automatic Indexing. Ph.D. Dissertation, The Ohio State University, Columbus, Ohio, in preparation.

8. A. Petrarca and W. M. Lay, The Double KWIC Coordinate Index. II: Use of an Automatically Generated Authority List to Eliminate Scattering Caused by Some Singular and Plural Main Index Terms, Computer and Information Science Research Center, The Ohio State University, Columbus, Ohio, 1969 (OSU-CISRC-TR-69-9).

9. D. S. Colombo and J. E. Rush, "Use of Word Fragments in Computer-Based Retrieval Systems," Journal of Chemical Documentation 9 (1), 47-50 (1969).

10. D. S. Colombo and J. E. Rush, "Use of Word Fragments in Computer-Based Retrieval Systems. II." Presented before the Division of Chemical Literature, American Chemical Society, 160th Meeting, Chicago, Illinois, September 1970.

11. P. L. Long, K. B. L. Rastogi, J. E. Rush and J. A. Wyckoff, "Large One-Line Files of Bibliographic Data" Presented before the IFIP Congress 71, August 23-28, 1971, Ljubljana, Yugoslavia.

12. C. L. Bernier, "Indexing Process Evaluation," American Documentation, 16 (4), 325-328 (1965).

13. Yehoshua Bar-Hillel, "The Present Status of Automatic Translation of Languages", in Franz L. Alt (ed.), Advances in Computers, Volume 1, Academic Press, New York, New York, 1960.

14. B. C. Landry and J. E. Rush, "Toward a Theory of Indexing--I" Proceedings of the American Society for Information Science, Volume 5, Greenwood Publishing Corporation, New York, New York, 1968.

15. B. C. Landry and J. E. Rush, "Toward a Theory of Indexing--II" Journal of the American Society for Information Science 21 (5), 358-367 (1970).

16. B. C. Landry and J. E. Rush, "Back-of-the Book Indexes and Indexing Theory", The Information Bazaar, Sixth Annual National Colloquium on Information Retrieval, Philadelphia, Pennsylvania, 1969.

17. E. R. Kolb, "Computer Printing Forecast for '70's", Datamation 16 (16), 29-31 (1970).

18. Optical Character Recognition and the Years Ahead, Proceedings of the International Business Forms Industry Conference, Hollywood, Florida, December 1968, Business Press, Elmhurst, Illinois, 1969.

19. E. Garfield, "Citation Indexes--New Paths to Scientific Knowledge" The Chemical Bulletin, 43 (4), 11 (1956).

20. E. Garfield, "Citation Analysis as a Tool in Journal Evaluation," Science 178 (4060), 471-479 (1972).

21. J. E. Armitage, and M. F. Lynch, "Articulation in the Generation of Subject Indexes by Computer", Presented before the Division of Chemical Literature, American Chemical Society, 153rd Meeting, Miami Beach, Florida, April 1967.

22. J. B. Fried, B. C. Landry, D. M. Liston, Jr., B. P. Price, R. C. Van Buskirk and D. M. Waschsberger, "Index Simulation Feasibility and Automatic Document Classification," Computer and Information Science Research Center, The Ohio State University, Columbus, Ohio, 1968 (OSU-CISRC-TR-68-4).

23. A. E. Petrarca and W. M. Lay, "The Double-KWIC Coordinate Index--A New Approach for Preparation of High-Quality Printed Indexes by Automatic Indexing Techniques," Journal of Chemical Documentation 9 (4), 256-261 (1969).

24. G. Salton, "Automatic Text Analysis," Science 168 (3937), 335-343 (1970).

25. J. E. Rush, "Work at CAS on Search Guides and Thesauri", Presented before the Chemical Abstracts Service Open Forum, Miami Beach, Florida, April 1967.

26. J. A. Williams, "Functions of a Man-Machine Interactive Information Retrieval System," Journal of the American Society for Information Science 22 (5), 311-317 (1971).

Supplementary Bibliography for Computer-Based Abstracting

In preparation for my research in computer-based abstracting, I examined many publications reporting the results of related research. I have cited many of these publications as direct references to portions of my dissertation. There are also many related publications which I have not cited directly. I have included a list of these publications here as a supplementary bibliography.

B- 1 ANZLOWAR BR
ABSTRACT AUTOMATION IN DRUG DOCUMENTATION.=
AUTOMATION AND SCIENTIFIC COMMUNICATION, H.P. LUHN (ED.),
AMERICAN DOCUMENTATION INSTITUTE, 26TH ANNUAL MEETING,
CHICAGO,ILLINOIS; OCTOBER 1963, 103-104.

B- 2 ATHERTON P YOVICH JC
STUDY OF PHYSICS ABSTRACTS -- ABSTRACTING THREE TYPES OF
JOURNALS.=
JOURNAL OF CHEMICAL DOCUMENTATION, 04(03), 157-163(1964).

B- 3 BARNETT MP FUTRELLE RP
SYNTACTIC ANALYSIS BY DIGITAL COMPUTER.=
COMMUNICATIONS OF THE ACM 05(10), 515-526(1962).

B- 4 BELKNAP RH
HOW TO IMPROVE SCIENTIFIC COMMUNICATION.=
JOURNAL OF CHEMICAL DOCUMENTATION, 02(03), 133-135(1962).

B- 5 BERNIER CL
CONDENSED TECHNICAL LITERATURES.=
JOURNAL OF CHEMICAL DOCUMENTATION, 08(04), 195-197(1968).

B- 6 BERZON VE
APPROACH TO PROBLEMS OF AUTOMATIC ABSTRACTING AND
AUTOMATIC TEXT CONTRACTION.=
NAUCHNO-TEKHNICHESKAYA INFORMATSIYA, SERIES 2,
INFORMATSIONNYE PROTSESSY I SISTEMY 1971(10) 16-21(1971).

B- 7 BOLDOVICI JA PAYNE D MCGILL OW
EVALUATION OF MACHINE-PRODUCED ABSTRACTS.=
TECHNICAL REPORT NO. RADC-TR-66-150, MAY 1966, ROME AIR
DEVELOPMENT CENTER, RESEARCH AND TECHNOLOGY DIVISION,
AIR FORCE SYSTEMS COMMAND, GRIFFISS AIR FORCE BASE,
NEW YORK, (AD 482 535).

B- 8 BOTTLE RT SCHWARZLANDER H
VARIATIONS IN THE ASSESSMENT OF THE INFORMATION CONTENT OF
DOCUMENTS.=
PROCEEDINGS OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE:
THE INFORMATION CONSCIOUS SOCIETY, JEANNE B. NORTH, (ED.),
VOLUME 7, 1970, PP279-281.

B- 9 BOURNE CP
THE WORLD'S TECHNICAL JOURNAL LITERATURE: AN ESTIMATE OF
VOLUME, ORIGIN, LANGUAGE, FIELD, INDEXING, AND
ABSTRACTING.=
AMERICAN DOCUMENTATION, 13(02), 159-168(1962).

B-10 BROER JW
ABSTRACTS IN BLOCK DIAGRAM FORM.=
IEEE TRANSACTIONS ON ENGINEERING WRITING AND SPEECH EWS-14(03), 64-67(1971)

B-11 CASEY RS
CHEMICAL INFORMATION: CONCEPTION AND PRENATAL CARE.=
JOURNAL OF CHEMICAL DOCUMENTATION, 02(03), 136-139(1962).

B-12 CHEMICAL ABSTRACTS SERVICE
REPORT ON THE TWELFTH CHEMICAL ABSTRACTS SERVICE OPEN FORUM.=
AMERICAN CHEMICAL SOCIETY, NEW YORK,NEW YORK, SEPTEMBER 7, 1969. CHEMICAL ABSTRACTS SERVICE, COLUMBUS,OHIO, JANUARY 1970.

B-13 CHEMICAL ABSTRACTS SERVICE
REPORT ON THE THIRTEENTH CHEMICAL ABSTRACTS SERVICE OPEN FORUM.=
AMERICAN CHEMICAL SOCIETY, TORONTO,CANADA, MAY 24, 1970. CHEMICAL ABSTRACTS SERVICE, COLUMBUS,OHIO, JANUARY 1971.

B-14 CHEMICAL ABSTRACTS SERVICE
REPORT ON THE FIFTEENTH CHEMICAL ABSTRACTS SERVICE OPEN FORUM.=
AMERICAN CHEMICAL SOCIETY, LOS ANGELES,CALIFORNIA, MARCH 30, 1971. CHEMICAL ABSTRACTS SERVICE, COLUMBUS,OHIO, JULY 1971.

B-15 CHEMICAL ABSTRACTS SERVICE
REPORT ON THE SIXTEENTH CHEMICAL ABSTRACTS SERVICE OPEN FORUM.=
AMERICAN CHEMICAL SOCIETY, WASHINGTON,D.C., SEPTEMBER 14,1971. CHEMICAL ABSTRACTS SERVICE, COLUMBUS,OHIO, JANUARY 1972.

B-16 COLLISON R
ABSTRACTS AND ABSTRACTING SERVICES.=
AMERICAN BIBLIOGRAPHIC CENTER - CLIO PRESS, SANTA BARBARA, CALIFORNIA, 1971.

B-17 DEBONS A OTTEN K
TOWARDS A METASCIENCE OF INFORMATION: INFORMATOLOGY.=
JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE, 21(01), 89-94(1970).

B-18 DEFENSE DOCUMENTATION CENTER
ABSTRACTING SCIENTIFIC AND TECHNICAL REPORTS OF DEFENSE SPONSORED RDT&E.=
DEFENSE DOCUMENTATION CENTER, CAMERON STATION, ALEXANDRIA, VIRGINIA, MARCH 1968, (AD 667 000)

B-19 DOYLE LB
THE MICROSTATISTICS OF TEXT.=
SYSTEM DEVELOPMENT CORPORATION, SANTA MONICA, CALIFORNIA, FEBRUARY 21, 1963, REPORT NO. SP 1083, (AD 401 445).

B-20 DOYLE LB
EXPANDING THE EDITING FUNCTION IN LANGUAGE DATA PROCESSING.=
COMMUNICATIONS OF THE ACM 08(04), 238-243(APRIL 1965).

B-21 DYSON GM RILEY EF
USE OF MACHINE METHODS AT CHEMICAL ABSTRACTS SERVICE.=
JOURNAL OF CHEMICAL DOCUMENTATION, 02(01), 19-22(1962).

B-22 EDMUNDSON HP
A STATISTICIAN'S VIEW OF LINGUISTIC MODELS AND LANGUAGE-DATA PROCESSING.=
PAUL L. GARVIN (ED.), NATURAL LANGUAGE AND THE COMPUTER, MCGRAW-HILL BOOK COMPANY, INC., NEW YORK, 1963, 151-179.

B-23 FLEISCHER M
SPEED OF ABSTRACTING AND COUNTRY OF ORIGIN OF ABSTRACTS PUBLISHED IN SECTION 8 (MINERALOGICAL AND GEOLOGICAL CHEMISTRY) OF CHEMICAL ABSTRACTS.=
JOURNAL OF CHEMICAL DOCUMENTATION, 01(02), 36-44(1961).

B-24 FRIEDENSTEIN H
ALERTING WITH INTERNAL ABSTRACT BULLETINS.=
JOURNAL OF CHEMICAL DOCUMENTATION, 05(03), 154-157(1965).

B-25 GARFIELD E GRANITO CE PETRARCA AE
INFORMATION RETRIEVAL SERVICES AND METHODS.=
ENCYCLOPEDIA OF CHEMICAL TECHNOLOGY, SUPPLEMENT VOLUME, 2ND ED., JOHN WILEY & SONS, INC., 1971, 510-535.

B-26 GARVIN PL
A LINGUIST'S VIEW OF LANGUAGE-DATA PROCESSING.=
PAUL L. GARVIN (ED.), NATURAL LANGUAGE AND THE COMPUTER, MCGRAW-HILL BOOK COMPANY, INC., NEW YORK, 1963, 109-127.

B-27 GARVIN PL
AN INFORMAL SURVEY OF MODERN LINGUISTICS.=
AMERICAN DOCUMENTATION, 16(04), 291-298(1965).

B-28 GOLDFARB CF MOSHER EJ PETERSON TI
AN ONLINE SYSTEM FOR INTEGRATED TEXT PROCESSING.=
THE INFORMATION CONSCIOUS SOCIETY, PROCEEDINGS OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE, VOLUME 7, 33RD ANNUAL MEETING, PHILADELPHIA, PENNSYLVANIA, OCTOBER 11-15, 1970.

B-29 GOLDFARB CF MOSHER EJ PETERSON TI
INTEGRATED TEXT PROCESSING FOR PUBLISHING AND INFORMATION RETRIEVAL.=
IBM DATA PROCESSING DIVISION, CAMBRIDGE SCIENTIFIC CENTER, G320-2065, APRIL, 1971.

B-30 GWIRTSMAN JJ
COVERAGE OF RUSSIAN CHEMICAL LITERATURE IN CHEMICAL ABSTRACTS.=
JOURNAL OF CHEMICAL DOCUMENTATION, 01(02), 38-44(1961)

B-31 HELMUTH NA
THE USE OF EXTRACTS IN INFORMATION SERVICES.=
JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE, 22(06), 382-389(1971).

B-32 HERNER S
SUBJECT SLANTING IN SCIENTIFIC ABSTRACTING PUBLICATIONS.=
PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON SCIENTIFIC INFORMATION, WASHINGTON,D.C., NOVEMBER 16-21,1958; NATIONAL ACADEMY OF SCIENCES - NATIONAL RESEARCH COUNCIL, WASHINGTON,D.C., 1959, 407-428.

B-33 HERNER S
EFFECT OF AUTOMATED INFORMATION RETRIEVAL SYSTEMS ON AUTHORS.=
AUTOMATION AND SCIENTIFIC COMMUNICATION, H.P. LUHN (ED.), AMERICAN DOCUMENTATION INSTITUTE, 26TH ANNUAL MEETING, CHICAGO,ILLINOIS, OCTOBER 1963, 101-102.

B-34 HOEGBERG EI
THE ABSTRACTOR AND THE INDEXER.=
JOURNAL OF CHEMICAL DOCUMENTATION, 02(03), 165-167(1962).

B-35 HOSHOVSKY AG
SUGGESTED CRITERIA FOR TITLES, ABSTRACTS, AND INDEX TERMS IN DOD TECHNICAL REPORTS.=
OFFICE OF AEROSPACE RESEARCH, WASHINGTON, D.C., OFFICE OF SCIENTIFIC AND TECHNICAL INFORMATION, OAR-65-9., OCTOBER 1965, (AO 622 944).

B-36 JACOBS RA ROSENBAUM PS
ENGLISH TRANSFORMATIONAL GRAMMAR.=
BLAISDELL PUBLISHING COMPANY, WALTHAM,MASS., 1968

B-37 KEENAN S
ABSTRACTING AND INDEXING SERVICES IN THE PHYSICAL SCIENCES.=
LIBRARY TRENDS, 16(03), 329-336(1968).

B-38 KEENAN S.
ABSTRACTING AND INDEXING SERVICES IN SCIENCE AND TECHNOLOGY.=
CA CUADRA (ED), ANNUAL REVIEW OF INFORMATION SCIENCE AND TECHNOLOGY, VOLUME 4, ENCYCLOPEDIA BRITANNICA,INC., CHICAGO, ILLINOIS, 1969, 273-303.

B-39 KEGAN DL
MEASURES OF THE USEFULNESS OF WRITTEN TECHNICAL INFORMATION TO CHEMICAL RESEARCHERS.=
JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE, 21(3), 179-186(1970).

B-40 KELLOGG C
THE FACT COMPILER: A SYSTEM FOR THE EXTRACTION, STORAGE, AND RETRIEVAL OF INFORMATION.=
PROCEEDINGS OF THE WESTERN JOINT COMPUTER CONFERENCE, SAN FRANCISCO, CAL., MAY 3-5,1960, VOL 17, P. 73-82.

B-41 KNABLE JP
AN EXPERIMENT COMPARING KEY WORDS FOUND IN INDEXES AND ABSTRACTS PREPARED BY HUMANS WITH THOSE IN TITLES.=
AMERICAN DOCUMENTATION, 16(02), 123-124(1965).

B-42 KNOWER BM
ABSTRACT JOURNALS AND BULLETINS.=
JOURNAL OF CHEMICAL DOCUMENTATION, 05(03), 150-153(1965).

B-43 KRAUSS RM
LANGUAGE AS A SYMBOLIC PROCESS IN COMMUNICATION — A PSYCHOLOGICAL PERSPECTIVE.=
AMERICAN SCIENTIST 56(3), 265-278(1968).

B-44 LANDRY BC RUSH JE
INDEXING AND ABSTRACTING. IV. AN INDEXING AND RE-INDEXING SIMULATION MODEL.=
PROCEEDINGS OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE, 33RD ANNUAL MEETING, PHILADELPHIA, PENNSYLVANIA, OCTOBER 11-15,1970, VOLUME 7, 269-274.

B-45 LAY WM PETRARCA AE
MODIFIED DOUBLE KWIC COORDINATE INDEX. REFINEMENTS IN MAIN TERM AND SUBORDINATE TERM SELECTION.=
THE COMPUTER AND INFORMATION SCIENCE RESEARCH CENTER, THE OHIO STATE UNIVERSITY, COLUMBUS,OHIO, 1970, TECHNICAL REPORT SERIES (OSU-CISRC-TR-70-10).

B-46 LEWENZ GF ZAREMBER GF BRENNER EH
STYLE AND SPEED IN PUBLISHING ABSTRACTS.=
JOURNAL OF CHEMICAL DOCUMENTATION, 01(02), 48-51(1961).

B-47 LIBBEY MA
THE USE OF SECOND ORDER DESCRIPTORS FOR DOCUMENT RETRIEVAL.=
AMERICAN DOCUMENTATION, 18(01), 10-20(1967).

B-48 LOWRY CD COCROFT R PASEK RL
ABSTRACTING SERVICES IN CLOSELY DEFINED FIELDS.=
JOURNAL OF CHEMICAL DOCUMENTATION, 06(04), 254-256(1966).

B-49 LYNN KC
A QUANTITATIVE COMPARISON OF CONVENTIONAL INFORMATION COMPRESSION TECHNIQUES IN DENTAL LITERATURE.=
AMERICAN DOCUMENTATION, 20(02), 149-151(1969).

B-50 MACKAY DM
THE INFORMATIONAL ANALYSIS OF QUESTIONS AND COMMANDS.=
INFORMATION THEORY, FOURTH LONDON SYMPOSIUM, C. CHERRY (ED.) LONDON, BUTTERWORTHS, 1961, 469-476.

B-51 MALONEY CJ
SEMANTIC INFORMATION.=
AMERICAN DOCUMENTATION, 13(03), 276-287(1962).

B-52 MARON ME KUHNS JL
ON RELEVANCE, PROBABILISTIC INDEXING AND INFORMATION RETRIEVAL.=
JOURNAL OF THE ACM, 07(03), 216-244(1960).

B-53 MARON ME
AUTOMATIC INDEXING: AN EXPERIMENTAL INQUIRY.=
JOURNAL OF THE ACM, 08(03), 404-417(1961).

B-54 MARON ME
A LOGICIAN'S VIEW OF LANGUAGE-DATA PROCESSING.=
PAUL L. GARVIN (ED.), NATURAL LANGUAGE AND THE COMPUTER, MCGRAW-HILL BOOK COMPANY, INC., NEW YORK, 1963, 128-150.

B-55 MOHLMAN JW
COSTS OF AN ABSTRACTING PROGRAM.=
JOURNAL OF CHEMICAL DOCUMENTATION, 01(02), 64-67(1961).

B-56 MOSHER EJ
INFORMATION SYSTEM ECONOMIES WITH A GENERALIZED MARKUP LANGUAGE.=
PRESENTED AT THE ASIS FIRST MID-YEAR REGIONAL CONFERENCE, MAY 19-21, 1972, THE UNIVERSITY OF DAYTON, DAYTON, OHIO

B-57 MURDOCK JW LISTON DM
A GENERAL MODEL OF INFORMATION TRANSFER: THEME PAPER 1968 ANNUAL CONVENTION.=
AMERICAN DOCUMENTATION, 18(04), 197-208(1967).

B-58 NOEL J
A SEMANTIC ANALYSIS OF ABSTRACTS AROUND AN EXPERIMENT ON MECHANIZED INDEXING.=
PH.D. DISSERTATION, UNIVERSITE DE LIEGE, 1972.

B-59 SALTON G
SOME EXPERIMENTS IN THE GENERATION OF WORD AND DOCUMENT ASSOCIATIONS.=
PROCEEDINGS OF THE 1962 FALL JOINT COMPUTER CONFERENCE, 234-250, 1962.

B-60 SALTON G
AUTOMATIC INFORMATION ORGANIZATION AND RETRIEVAL.=
MCGRAW-HILL BOOK COMPANY, NEW YORK, N.Y., 1968

B-61 SALTON G
A COMPARISON BETWEEN MANUAL AND AUTOMATIC INDEXING METHODS.=
AMERICAN DOCUMENTATION, 20(01), 61-71(1969).

B-62 SASTRI MI
A NOTE ON THE STYLE OF CHEMICAL ABSTRACTS.=
JOURNAL OF CHEMICAL DOCUMENTATION, 06(02), 71-72(1966).

B-63 SCHREIDER YA
SEMANTIC ASPECTS OF INFORMATION THEORY.=
INTERNATIONAL FEDERATION FOR DOCUMENTATION STUDY COMMI3355 <RESEARCH ON THEORETICAL BASIS OF INFORMATION> (FIO/RI), FID 435, ALL-UNION INSTITUTE FOR SCIENTIC AND TECHNICAL INFORMATION, MOSCOW, 1969.

B-64 SCHREIDER YA
ON THE SEMANTIC CHARACTERISTICS OF INFORMATION.=
INFORMATION STORAGE AND RETRIEVAL 2(4), 221-233(1965).

B-65 SCHULTZ L
REESTABLISHING THE DIRECT INTERFACE BETWEEN AN ABSTRACTING AND INDEXING SERVICE AND THE GENERATOR OF THE LITERATURE.=
PROCEEDINGS OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE, ANNUAL MEETING, VOLUME 5, INFORMATION TRANSFER, COLUMBUS, OHIO, OCTOBER 20-24, 1968, GREENWOOD PUBLISHING CORPORATION, NEW YORK, 1968.

B-66 SHANNON CE WEAVER W
THE MATHEMATICAL THEORY OF COMMUNICATION.=
THE UNIVERSITY OF ILLINOIS PRESS, URBANA,ILLINOIS, 1964

B-67 SLAMECKA V ZUNDE P
AUTOMATIC SUBJECT INDEXING FROM TEXTUAL CONDENSATIONS.=
AUTOMATION AND SCIENTIFIC COMMUNICATION, H.P. LUHN (ED.), AMERICAN DOCUMENTATION INSTITUTE, 26TH ANNUAL MEETING, CHICAGO,ILLINOIS, OCTOBER 1963, 139-140.

B-68 SMITH MH
AN EVALUATION OF ABSTRACTING JOURNALS AND INDEXES.=
PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON SCIENTIFIC INFORMATION, WASHINGTON,D.C., NOVEMBER 16-21,1958; NATIONAL ACADEMY OF SCIENCES - NATIONAL RESEARCH COUNCIL, WASHINGTON,D.C., 1959, 321-350.

B-69 STILES HE
THE ASSOCIATION FACTOR IN INFORMATION RETRIEVAL.=
JOURNAL OF THE ACM, 08(02), 271-279(1961).

B-70 SWANSON DR
SEARCHING NATURAL LANGUAGE TEXT BY COMPUTER.=
SCIENCE, 132(3434), 1099-1104(1960).

B-71 THOMPSON P
TITLES AND ABSTRACTS: A GUIDE TO THEIR PREPARATION FOR ENGINERRING AUTHORS.=
TECHNICAL INFORMATION SERIES (R70ELS-51), GENERAL ELECTRIC, AUGUST 1970.

B-72 WALTER GO
TYPESETTING.=
SCIENTIFIC AMERICAN 220(05), 60-69(1969).

B-73 WEIL BH
WRITING THE LITERATURE SUMMARY.=
B.H. WEIL (ED.), THE TECHNICAL REPORT, REINHOLD PUBLISHING CORPORATION, NEW YORK, 1954.

B-74 WEIL BH
SOME READER REACTIONS TO ABSTRACT BULLETIN STYLE.=
JOURNAL OF CHEMICAL DOCUMENTATION, 01(02), 52-58(1961).

B-75 WEIL BH ZAREMBER I OWEN H
TECHNICAL-ABSTRACTING FUNDAMENTALS. I. INTRODUCTION.=
JOURNAL OF CHEMICAL DOCUMENTATION, 03(02), 86-89(1963).

B-76 WEIL BH ZAREMBER I OWEN H
TECHNICAL-ABSTRACTING FUNDAMENTALS. II. WRITING PRINCIPLES AND PRACTICES.=
JOURNAL OF CHEMICAL DOCUMENTATION, 03(03), 125-132(1963).

B-77 WEIL BH ZAREMBER I OWEN H
TECHNICAL-ABSTRACTING FUNDAMENTALS. III. PUBLISHING ABSTRACTS IN PRIMARY JOURNALS.=
JOURNAL OF CHEMICAL DOCUMENTATION, 03(03), 132-136(1963).

B-78 WEIL BH
STANDARDS FOR WRITING ABSTRACTS -- AUTHOR'S REPLY.=
JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE, 22(03), 229(1971).

B-79 WELLISCH H
STANDARDS FOR WRITING ABSTRACTS -- LETTER TO JASIS EDITOR.=
JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE, 22(03), 227-229(1971).

B-80 WHITTEMORE BJ
A GENERALIZED DECISION MODEL FOR THE ANALYSIS OF INFORMATION.=
PH.D. DISSERTATION, THE OHIO STATE UNIVERSITY, COLUMBUS, OHIO, AUGUST 1972.

B-81 YOVITS MC ERNST RL
GENERALIZED INFORMATION SYSTEMS: SOME CONSEQUENCES FOR INFORMATION TRANSFER.=
TECHNICAL REPORT NO.68-1, COMPUTER AND INFORMATION SCIENCE RESEARCH CENTER, THE OHIO STATE UNIVERSITY, COLUMBUS, OHIO, OCTOBER,1968.

KWIC Index of References in Computer-Based Abstracting

The following pages contain a KWIC (Key-Word-In-Context) index of the cumulated references from the preceeding chapters and the supplementary bibliography. Using a stop list of 81 common and function words, 185 bibliographic records generated 1055 index entries. The index entries are derived from significant words from the titles and from the authors' last names.

The four character accession number that accompanies each entry serves to indicate the citation of each reference. The number may be of the following form:

2-24

where the digit preceding the hyphen indicates the chapter number and the two digits following the hyphen indicate the reference number within the chapter. If more than one accession number is listed for a document, this indicates that the document is cited in more than one location in the dissertation. The reader can find an indication of the content of the document by finding the portion of text that references it. The accession number may also be of the following alternate form:

B-53

where B indicates that the reference is included in the supplementary bibliography and the two digits following the hyphen indicate the number assigned to the reference.