Automatic language processing under the plan includes mechanical translation, computational linguistics, and related work in areas such as automatic abstracting and development of hardware. The adoption of this plan constitutes recognition by the agencies involved that "fully automatic high-quality language processing, including mechanical translation, is a long-range goal," and that cooperation in planning and research are necessary to progress in this field.

A recent important step under the Joint Automatic Language Processing Program has been the appointment by the National Academy of Sciences of John R. Pierce as chairman of an advisory committee for automatic language processing. When fully established, the committee will provide the agencies participating in the joint program with advice which will aid them in planning future research, development, and evaluation in this field.

Warren Weaver concludes his note "Translation," referred to earlier (*3*), by indicating four possible types of attack on the mechanical translation problem. As for the fourth, he says,

> Indeed, what seems . . . to be the most promising approach of all is one based on . . . an approach that goes so deeply into the structure of languages as to come down to the level where they exhibit common traits. . . .
>
> Such a program involves a presumably tremendous amount of work in the logical structure of languages before one would be ready for any mechanization. . . . But it is along such general lines that it seems likely that the problem of translation can be attacked successfully. Such a program has the advantage that, whether or not it lead to a useful mechanization of the translation problem, it could not fail to shed much useful light on the general problem of communication.

The history of mechanical translation research has shown Warren Weaver's insight of 15 years ago to have been remarkably prophetic.

References and Notes

1. J. B. Wiesner, "Communication sciences in a university environment," paper presented at the Conference on Scientific Information, San Jose, California, 1958, and published in *IBM J. Res. Develop.* **2**, 271 (1958).
2. Y. Bar-Hillel, "The present status of automatic translation of languages," in *Advances in Computers*, F. L. Alt, Ed. (Academic Press, New York, 1960), vol. 1, p. 135.
3. The entire note is reproduced in *Machine Translation of Languages*, W. N. Locke and A. D. Booth, Eds. (Massachusetts Institute of Technology Press, Cambridge, 1955), pp. 15–23.
4. *Sci. Inform. Notes* **5**, No. 6 and 7 (1963–1964).
5. W. D. Climenson, N. H. Hardwick, S. N. Jacobson, *Am. Doc.* **12**, No. 3, 178 (1961).
6. G. Salton, *ibid.* **14**, No. 3, 213 (1963).
7. I am grateful to the National Bureau of Standards and to the project leader, Mrs. Ida Rhodes, for the opportunitiy to participate for a time in research there on mechanical translation.
8. C. Dougherty, S. M. Lamb, S. E. Martin, "Chinese Character Indexes" (Univ. of California Press, Berkeley, 1963).
9. The Bunker Ramo Corporation was formerly the Ramo-Wooldridge Division of Thompson Ramo Wooldridge, Inc.
10. Y. Bar-Hillel, "Machine translation: The end of an illusion," in *Information Processing 1962* (proceedings of the International Federation for Information Processing Congress, held in Munich, 1962), C. M. Popplewell, Ed. (North-Holland, Amsterdam, 1963), p. 331.

Automatic Information Processing in Western Europe

Current European work in automatic documentation and information processing is reviewed and evaluated.

Gerard Salton

It is well known that European work in the design and applications of computing equipment is widely underrated in the United States. Few Americans read the foreign technical journals, and the assumption is widespread that European work is either inferior in quality or, in any case, lagging behind equivalent American work by many years.

This opinion seems astonishing when one considers, for example, the outstanding European contributions in the area of computer organization and computer language design. However, many Europeans appear to agree with the view that one must come to America in order to be in the forefront of developments. The following quotation from Fairthorne may be typical (*1*):

> In the English system, in my experience, one first pretends that the [information processing] log jam does not exist. Then, when it is absolutely impossible to avoid knowing it is there, one says that, after all, only the very best people do have log jams and this is good for the national character.

Conditions in Europe were found to be neither all black nor all white. Clearly, in quantity at least, the European output cannot compare with the work in this country. On the other hand, a considerable number of interesting developments are taking place.

In this article an attempt is made first to describe briefly the conditions of research in Europe and, thereafter, to cover some of the more interesting recent endeavors in information processing. Developments in processor and information organization are mentioned briefly, followed by a selective review of work in the analysis of natural languages; the identification of document content by statistical, structural, and logical methods; the generation of computer-produced outputs of various kinds; the operation of automatic information retrieval systems; and the evaluation of retrieval effectiveness (*2*).

Conditions of Research

It must be mentioned at the outset that the importance of automatic information processing is being increasingly realized in Europe; this is directly reflected, for example, in the rapid pace at which computers are being installed

The author is assistant professor of applied mathematics at the Computation Laboratory of Harvard University, Cambridge, Mass. This article was prepared while he was a Guggenheim fellow at the I.B.M. Research Laboratory, Zurich, Switzerland, from February to September 1963.

(*3*). However, computing equipment is still relatively scarce, and where it *is* available, priority is ordinarily given to the solution of numeric rather than non-numeric problems. As a result, the emphasis in information processing is primarily on theoretical work, rather than on computer experiments, which often are expensive and may or may not pay off.

Two main characteristics of European work become evident when one attempts to draw a comparison with American conditions. First, many of the European groups demonstrate a good deal of imagination both in the approach which they choose to take and in the solutions which they eventually propose for a given problem, and while imagination by itself is usually not enough, it is nevertheless an important ingredient of a useful research program. It may be that some of the more unusual ideas are due to the broader background of many of the European workers (the narrow specialization common in this country is relatively rare in Europe), or possibly it is simply that many European groups have a freer hand, not being hampered by the kind of directed pressure which often results from certain narrowly interpreted contractual obligations. In any case, in an interdisciplinary field which lacks an accepted theory, an imaginative approach would seem to be a real asset.

The second main impression one gets of conditions in Europe is a less positive one: there is a great deal of unnecessarily conservative thinking in matters of research. The curriculum in the computer sciences changes much more slowly at most European universities than it does in this country, and many students are not afforded an adequate introduction into the newer areas of non-numeric computer applications. At the same time, it is often difficult to obtain adequate funding for either theoretical or experimental work, and one sometimes receives the impression that nothing less than a written guarantee of ultimate success will render a given project acceptable to many of the fund-granting agencies. Furthermore, proposals for research are often not judged on scientific merit alone, and whether or not a given piece of work receives support depends, at least in part, on how well it fits in with the national aims and policies of a given country. This situation is especially acute and detrimental in some of the inter-European organizations supervised by multinational committees. Such organizations frequently have their own research branches, and function also as fund-granting agencies for outside groups. Since technical decisions are made by international committees with diverse views and aims, the projects which receive support are sometimes those which are innocuous enough not to offend any of the participating nationalities, rather than those which are really worthy of being pushed.

To summarize, if there is any lag in European output in the information-processing area it is a lag in quantity, compounded by a continuing scarcity of first-rate computing equipment in many cases, and by certain unnecessarily restrictive research policies. Some highly original and worthwhile contributions are, however, in evidence, and developments in information processing may be expected to accelerate as computing facilities become more generally available and research policies are modified.

Some of the information-processing research is described in more detail in the next sections.

Processor Organization

The European contributions in the area of processor organization are so well known that only brief mention is made here of some of the most important developments. It has been known for some years that a simple storage mechanism operating on a last-in-first-out basis, and known as the "pushdown store" (*4*), could be used to advantage for analyzing and evaluating statements in certain semiformal languages. So far as I know, the first proposed design of a machine incorporating a pushdown store is in a German patent (*5*), and work in which pushdown algorithms are used for translating a variety of processing languages has been pursued both here and abroad (*6, 7*). Moreover, at least two computers incorporating physically implemented pushdown stores have actually been built (*8*). It is not unlikely that the continued development in Europe of formally defined processing languages, such as Algol, is influenced directly by the simplicity of the pushdown principle as it affects the analysis and translation of these languages.

Interest in parallel computation is also widespread in Europe as attested by the construction in France of the Gamma-60 computer (*9*) and in England of the Atlas (*10*). The former, which has a number of independent processing units, is a precursor of the parallel-access, cell-type computers which have received considerable attention in recent years in this country (*11*), and the latter is of great interest in information processing because of its effect on time-sharing operations.

The Altas computer has a hierarchy of auxiliary memory structures, with relatively slow-access mechanisms, tied to a principal high-speed memory. The high-speed memory must contain both the data and the instructions which are to be operated upon by the computer at any given moment. However, to the user this restriction is of no consequence because a supervisory program is available which automatically handles the transfer of data from slow to fast storage and vice versa. The complete hierarchy of different memories thus appears, for all practical purposes, as a single, large, "one-level" store.

The main high-speed memory is separated into blocks, known as "pages," and one whole page is transferred at a time from one store to the other. "Page-turning" algorithms are provided for choosing in a reasonable manner the particular page which must be removed from the high-speed store in order to make room for a new page. The various pages in the memory may contain a number of different programs, belonging either to a single user or to several distinct users, and the page-turning facilities can be used to switch from one program to another. The Atlas thus lends itself to the type of operation in which a number of different customers use the same central machine, and do so almost simultaneously.

This concept, supplemented by a priority system to determine the order in which different users are given access to the equipment and by trapping features to interrupt a user's program when certain predetermined conditions arise, is of fundamental importance in information processing, since any operating automatic information-retrieval system may be expected eventually to operate as a time-sharing system, accommodating many different users at the same time. Time-sharing systems are receiving increasing attention, both in this country (as evidenced, for example, by the work on project MAC

at the Massachusetts Institute of Technology) and abroad (*12*).

Another European development in processor organization which should be mentioned is the implementation of a linkage system to tie together several different digital and analog computers (*13*). It is hoped that such a hybrid system might be used to advantage by having the digital part supply the control variables for the analog machines and by using the analog equipment as additional processors, supplementing the digital devices. This type of organization raises interesting engineering and programming problems, on which work is in progress.

Processing Languages and Information Organization

A large part of the development and maintenance work on Algol and Algol-like languages has been concentrated in Europe for several years. Moreover, there exists a continuing high interest in the construction of improved programming languages and the production of improved translation systems from one language to another. Whereas in this country much work is directed toward improving processing performance through the development of clever storage arrangements and fast search and retrieval techniques, in Europe the emphasis seems to be on fundamental work in programming languages, both of a theoretical and of a more practical kind (*14*). Algol compilers have been prepared for a large number of machines, and considerable effort has been devoted to the construction of versatile, syntax-directed compiling systems (*7, 15*). New programming languages are also being generated, as are extensions of existing ones (*16, 17*). However, relatively little work is being done to determine the structure of the information to be processed and to use it in the generation of processing algorithms. An exception is the inclusion in some of the programming languages of special operations believed to be useful for the manipulation of special constructs. Thus, a study of learning processes has led to the construction of an Algol-like language, useful for the transformation of logical vectors and matrices (*17, 18*). Similarly, a programming language has been designed for the manipulation of graphs (*19*), and macro-operations have been proposed for use in automatic documentation systems (*20*).

Another interesting development is the design of a new language, called Lotis, to describe the structure (including logic, timing, and sequencing) of digital machines. This language makes it possible to simulate the operations of a machine, given its design, and should become particularly useful during the construction of multiprocessing and time-sharing systems. Potential concurrency of operations can be expressed through partitioning the set of operations into "groups," each group containing the subset of steps generated by a particular control mechanism of the machine. Since any step in a sequence may activate one or more sequences in different groups, and because provisions are made for conditional entry into sequences, it is easy to represent arbitrarily complex time-sharing systems. Asynchronous, fixed-delay, or synchronous timing, or any combination of these, may be represented, and the timing specifications may be implicit or explicit (*21*).

Interest in list processing and in the construction of complex storage maps is not as yet very widespread in Europe, probably because there has been no opportunity so far to operate with large masses of structured data. I know of at least one list-type arrangement that has been proposed for the storage of language data (*22*), and further work in this area is likely, as more practical experience is gained in information processing (*23*).

Natural-Language Analysis

As might have been expected, a great deal of work in language analysis is being done in Europe. Since it is impossible to cover all aspects of this work, the present discussion is restricted to efforts in which use has actually been made of computing equipment at some point in the analysis. In this section I describe work primarily oriented toward the automatic translation of languages. In later sections I discuss linguistic work that deals more specifically with the analysis of document content for use in information retrieval.

Quite a few efforts are directed toward the morphological analysis of languages. Specifically, it is argued that for inflected languages, such as German and French, it is impractical to store in an automatic dictionary all possible word forms which may be constructed from a given word stem. Accordingly, a word occurring in a written text must be reduced to a standard canonical form before it can be located in the dictionary. Reduction algorithms have been generated for German and Russian (*24*), and more recently for French (*25*).

In the latter case an IBM 1401 was used to obtain the desired canonical forms, by first taking each word ending and applying transformation rules to generate a number of "hypothetical" canonical forms. Each hypothetical form might, in theory, have given rise to the original inflected form. In order to avoid a requirement for a very large number of transformation rules to handle all possible word forms in the language, a dictionary of exceptions is used which furnishes the canonical forms for a number of troublesome cases. A general dictionary is used as the last step in the procedure, to eliminate all those hypothetical forms which do not in fact occur in the language. The remaining forms are then accepted as correct. This relatively simple procedure has been incorporated in an automatic dictionary system used as an aid to human translators within the European Economic Community (EEC), and the system is apparently operating effectively (*26*).

Considerable attention has also been given to the syntactic analysis of natural languages, and in particular to a type of "dependency" analysis where each word is considered an operator with or without arguments (*27*). A syntactically analyzed sentence can be represented in such a system by a "dependency tree" whose branches identify the syntactic dependencies between the words (*28*). An interesting extension of the dependency analysis has been proposed recently to include punctuation marks on the same basis as ordinary words. This system establishes a criterion of grammaticality for sentences which is largely independent of the punctuation (*29*).

Two properties are defined first for punctuation marks: the separating power and the syntactic function. The former is independent of the grammatical environment and depends only on the given sign, and the latter is a function of the syntactic role of the punctuation. A scale of values is then defined, to measure the strength of syntactic connections between words in a sentence (the article-noun link being the strongest, followed by subject-verb, and so on), and each punctuation mark is assigned one of the syntactic-

connection values, as a function of its syntactic role in the sentence. Separating-power and syntactic-connection values are now used to state "coherence" laws which must be obeyed by any sequence of punctuation marks. It is found, in particular, that the separating power of two adjacent punctuation marks cannot increase if the syntactic-connection value increases, and vice versa. Typical "trajectories" are drawn for the sequence of punctuation marks in a sentence, by using the pairs of coefficients attached to the marks as Cartesian coordinates in a Euclidian plane.

Another language-analysis program which should be mentioned is the one based on the so-called "operational" method (*30*). The fundamental idea is that, in order to analyze a given sentence, it is necessary to concentrate on the elements of thought which are expressed. These elements of thought are represented by individual elements of meaning, rather than by individual words, and a complete thought process is then mirrored in the language by a correct correlation or juxtaposition of the various elements of meaning into larger and larger units, until finally a single "correlational net" is obtained for the whole sentence.

Specifically, various elements are first recognized in a sentence; these may be individual words, or particles such as prefixes or suffixes. By means of a dictionary "look-up" procedure each element is provided with a set of indicator pairs. The first member of each pair represents the type of correlation into which the given unit can enter, and the second member represents the place in the correlation which the given unit may occupy (each correlation consists of two correlates and a correlator, and a given unit may function either as a correlate or as the correlator, or as both correlator and correlate). Since a given unit of meaning can obviously be a part of many different types of correlations, a large number of indicator pairs are normally applicable in each case.

It is now necessary to use rules which permit the construction of large correlational nets from smaller ones. In particular, in order to be correctly combined, two elements must be of the same type—that is, must have identical first indicators—and must fit into position within the correlational net. All correlations which do not obey these restrictions are rejected, and equivalent correlational nets are eliminated in accordance with certain rules. The expectation is that at the end of the process only a single well-formed correlational net will have been generated for each nonambiguous sentence. The procedure is reminiscent of the so-called "morsel" technique which has been used with some measure of success for the automatic syntactic analysis of texts. Of course, the "operational" method can be successfully carried out only if detailed analyses of all possible correlations applicable to all the particles in the language can be produced. Whether such a complete semantic analysis can in fact be generated remains to be seen. An example worked out in detail gives a good account of the procedure (*31*).

Automatic Classification by Statistical Techniques

The simplest model for the representation of information is one where each item is identified by a set of independent properties. Such a system can be represented by an item-property matrix in which the ijth matrix position is set equal to 1 (or to some other numeric value different from zero) if the jth property applies to the ith item; otherwise, the ijth matrix position is set equal to zero. Given such an item-property matrix, it is possible to generate an item-item similarity matrix, by using as a criterion of similarity some function of the number of properties which are jointly assigned to two given items. Similarities between pairs of properties are derived in the same way by using as an index of similarity the number of items which have both properties included in their property sets.

In an information-retrieval system the items of information might be documents and the properties could be key words assigned to the documents. Quantitative procedures designed to measure similarities between either key words or documents could then be used in practice to supplement document descriptions by including new key words similar to the ones originally assigned; alternatively, document sets obtained, for example, in answer to search requests can be supplemented by the inclusion of other, related documents.

Instead of computing similarities only between pairs of items, it is possible to generate "clusters" of items, such that all items within a cluster are related more strongly to each other than to any item outside the cluster. New, unknown items can then be classified automatically through assignment to the cluster of items which exhibits the most nearly matching property set.

In the United States, extensive work directed toward the generation of effective correlation coefficients between items, and toward the use of term and document associations in automatic information classification and retrieval has been carried out. This work is not matched by a comparable effort in Europe. The most obvious explanation is, of course, the lack of accessible computing equipment to perform the experimental work. In addition, in Europe there is a deep distrust of the statistical methodology for the analysis of information. It is pointed out that properties identifying items of information are not generally independent but exhibit a variety of relations which are usually disregarded in the quantitative model. Furthermore, problems arise when it becomes necessary to explain the meaningfulness of some of the correlation experiments.

Despite these misgivings, some work has been done in these areas by various European groups. The work on clustering, or clumping, performed in England over the past few years is quite well known (*32*). Comparisons have been made, in particular, between automatic and manual assignment of unknown items to existing item clusters, and between clusters derived automatically and others generated by various manual procedures (*33*).

Some interesting experiments have also been performed recently in which term clusters were generated by permutation of a term-term similarity matrix so as to group the "ones" of the matrix in boxes along the diagonal (*34*). This form of the similarity matrix visually exhibits the item clusters, since each submatrix of "ones" then represents one of the clusters.

Thesaurus Techniques and Structural Analysis

When written documents are analyzed, the properties most immediately available for information identification are those based on the words which occur in the documents. The statistical techniques described in the preceding section would not, however, be very effective if they were applied to the

complete, unmodified vocabulary of the original texts. As a result, a number of refinements are normally introduced in the analysis; these consist of the use of synonym dictionaries or thesauruses to control the vocabulary, and of structural techniques to determine relations between words. The actual analysis is then based on a controlled vocabulary in which at least some relations between the properties are recognized.

A number of experiments have been performed in which statistical procedures for the generation of document identifiers (key words) are used in conjunction with synonym dictionaries. One generates such key words by first performing a thesaurus look-up and then picking as key words all those thesaurus headings which occur more than the expected number of times (*35*). Other recent work in vocabulary control has included a comparison between a manually produced index and a set of key words generated automatically with the help of a thesaurus. The terms included in the manual index are used to improve the thesaurus, which serves to normalize the vocabulary of the original documents (*36*). Thesauruses and key-word lists of various types, automatically or manually produced and with or without cross references between entries, are also widely used as a part of many conventional information systems (*37*).

The construction of several structural models for the representation of information is probably the most noteworthy European contribution in the information-retrieval area. In such models, provision is made for representing not only the individual properties of a given item but also a variety of relations between the properties. An item of information, such as a document, may be represented in such a system by a two-dimensional diagram, or graph, where the nodes and the branches of the graph stand, respectively for the properties and the relations between properties. To identify the relations automatically, a combination of syntactic, semantic, and logical analyses is usually required.

One of the best-known of the graphical models for use in documentation systems is the "general diagram" developed by Euratom (*38*). The general diagram is a graph designed to represent a given field of knowledge; it consists of nodes, which stand for "properties" and "objects," and of directed branches, which represent "actions" and "relations." Properties and objects are normally represented in the natural language by nouns and adjectives, while actions and relations are indicated by verbs and prepositions, respectively. At least 18 principal types of relations are recognized, and there are many subdivisions within the relational types. Provision is also made for representing in the general diagram coordinating information normally conveyed by conjunctions and similar particles. A given document is then identified through location of its "partial" diagram within the general diagram, and information is retrieved, in response to search requests, through comparison of the respective partial diagrams.

The general diagram is rather attractive as a model for the representation of information content, because much of the actual information is preserved in the diagram. On the other hand, it is difficult enough to construct a partial diagram for a given document manually, and to produce one automatically should be much harder. There arises, therefore, a question about the practical implementation of the system.

A somewhat similar model, similar in concept at least, is the Syntol system developed in France (*39*). In that system, document identifications are again represented as abstract graphs, and relations between the various identifiers are denoted by directed branches between the nodes. However, only four kinds of relations are recognized, and a given Syntol graph is in general easier to generate than the corresponding diagram in the system just described. Moreover, quite detailed analyses have been carried out, and there is a real chance that automatic implementation may eventually be achieved.

In the Syntol system a text is first segmented into individual words. Each word is then automatically looked up in a dictionary, and for certain (but not all) words, one obtains (i) the formal category (predicate *P*, entity *E*, state *S*, or action *A*); (ii) the corresponding Syntol term; and (iii) the semantic class. The Syntol terms are names of concepts which appear as nodes in the Syntol graph, and the semantic classes are indicators which place the various terms within a hierarchical organization constructed for the given field. The hierarchy effectively functions like a thesaurus and permits replacement of Syntol terms by more general notions or by more refined ones.

Four different relations are recognized: a formal "coordinative" relation, a dynamic "consecutive" relation to signify temporal succession, and two static relations, called, respectively, the "associative" and the "predicative." To obtain these relations, a type of syntactic analysis is used which involves both formal and semantic procedures. Specifically, so-called syntactic "tools" are recognized; these tools are represented by individual words or by sets of words in the language, and are classified as either strong or weak. A "grid" is then formed by a sequence of interdependent tools, and a syntactic construction is determined in Syntol as a function of the given syntactic grid and the formal or semantic classes of the surrounding Syntol words. For example, if the grid were, "effect of . . . on . . .," with the blanks filled by an action (*A*) and an entity (*E*), respectively, the corresponding Syntol graph would include an "*A*" node and an "*E*" node, with a relation of type 3 pointing from *A* to *E*. A number of formal rules, the so-called automatic "developments," also serve for the construction of syntactic connections, in addition to the rules based on grid types and surrounding Syntol categories. Thus, a construction $E \rightarrow S \rightarrow A$ also implies a construction $E \rightarrow A$; similarly, $E \rightarrow S \leftarrow A$ is developed into $A \rightarrow E$, and so on.

Among the most interesting aspects of the system are the changes and variations which make it possible to modify both requests for information and matching criteria between requests and stored information so as to produce varying amounts of information in response to the search requests. Thus, it is possible to use the hierarchical arrangement to modify certain Syntol terms, and to alter the matching conditions by disregarding or changing the syntactic relations between terms. In theory, this should make it possible to generate a set of rules by which a given user might specify a sequence of processing alterations to produce exactly the desired amount of information.

It remains to be seen whether the large number of syntactic rules which generate the relational indications can really be constructed, and whether the automatic detection of syntactic grids will operate satisfactorily for most

grids. Questions also arise about the practicability of the request alterations and the processing changes. In general, however, the development of Syntol is a promising sign of progress in information processing.

Computer-Produced Output and Operating Systems

It is in the area of computer-produced outputs and experimental operating systems that the differences between conditions in Europe and in the United States become most noticeable. There do exist, of course, in almost every European country, centers which use punched-card or data-processing equipment for the processing of language data (*40*). However, much of the work done is standard in the sense that only well-known techniques are used. There are exceptions, and a few of these more noteworthy efforts are outlined in this section.

Consider, first, the automatic production of bibliographies, indexes, collections of abstracts, and catalogues. This type of work is of increasing interest in Europe, as it is elsewhere, and a great deal of attention is devoted to the mechanization of a variety of indexing and abstracting services. One interesting development is the generation of a "key word in context" (KWIC) index, modified through deletion, from the permuted titles, of information considered to be of secondary importance (*41*). A simple type of syntactic analysis is used to isolate so-called "separator" words, of which some are compulsory and some are optional. These separator words are then used to isolate significant words and phrases, and the significant words and phrases are permuted by the usual KWIC process. The result is a printed index which is easier to scan than the normal KWIC index, yet preserves all the essential advantages, such as speed of production and multiple entry points.

Another interesting service in the "current awareness" category is one in which incoming documents are abstracted by human abstractors and classified into 17 major subject categories. Abstracts are then printed four to a page and duplicated, and individual booklets of abstracts are distributed to the participating scientists in accordance with the recipients' interests (*42*). The service is thus a simple type of "selective dissemination of information." When the documents are abstracted, sets of key words are assigned; some of these are from a list of controlled terms, and some are freely chosen terms. These key words are then used in a retrieval system programmed for a 1401 computer (*43*)

A few other operating systems deserve mention, among them a punched-card system in which a clever arrangement of superimposed codes is used for storing key-word information (*44*); a system for the retrieval of chemical structures, based on use of an automatic scanning device to introduce chemical structures into the system (*45*); and a search system in which the simplest kind of punched-card sorting and listing operations are used, for detecting possible trademark infringements (*46*). This last system is a fine example of the intelligent use of very simple procedures to accomplish relatively complicated tasks, such as the detection of a similarity in appearance or in represented sounds between two given sequences of letters representing trademarks.

Evaluation of Retrieval Systems

The evaluation of complex systems is always a matter of considerable difficulty, particularly since human judgment is in general required at some point in the process. Furthermore, even if the subjective factor were somehow to be eliminated, still there exists no generally effective way of determining which part of the system is in fact being evaluated. In the information-retrieval area, for example, attempts to compare the value of indexing systems may turn out to measure, instead, the accuracy of the human indexer who originally assigns the terms to the documents, or the effectiveness of the matching system which compares information requests with stored information.

In England, several studies have been made of the relative effectiveness of a variety of document-indexing systems, and elaborate testing systems have been designed to eliminate the influence of extraneous factors (*47*). In the most recent model the recall-precision ratio is used as a parameter for evaluating a retrieval system (*recall* is defined as the number of retrieved-and-relevant documents divided by the number of relevant documents in the collection; *precision*, on the other hand, is the number of retrieved-and-relevant documents divided by the total number of documents retrieved). Various procedures are then defined for widening the coverage of the index terms (for example, by including synonyms and by adding statistically associated terms) or, alternatively, for restricting the coverage by using relations between terms. The former procedure might be expected to improve recall, thus increasing the recall-precision ratio, while the latter procedure should have the reverse effect. By comparing the retrieval performance under a variety of conditions, some measure of the power of the various indexing procedures should be obtainable.

Studies of this nature are particularly timely because of the present interest in systems which permit interaction between the user and the system. In such semiautomatic systems the human investigator would be given directions by the computer concerning the steps which would most effectively increase or decrease recall at any given time. This flexible kind of systems organization is being actively considered in Europe as well as in the United States.

References and Notes

1. R. A. Fairthorne, "An outsider inside information: USA 1961–1962," *Aslib Proc.* **14**, 380 (1962).
2. For supplementary material, see R. A. Fairthorne (*1*); C. D. Gull, "Automatic documentation, current systems and trends in the USA," *Rev. Intern. Doc.* **29**, No. 2, 57 (1962); G. J. Koelewijn, "Recent developments in Western Europe in the field of the automation of document retrieval systems," *ibid.*, p. 42.
3. It has been estimated that by 1970 more than 10,000 computers will be in operation in the six European Economic Community countries alone [W. K. de Bruijn, A. B. Frielink, B. Scheepmaker, "Development of the Computer Market in Western Europe," *Netherlands Automatic Information Processing Research Center, Amsterdam, Rept.* (June 1963)].
4. Pushdown stores are also known as stacks or information cellars.
5. F. L. Bauer and K. Samelson, German patent DAS 1094019 (Dec. 1960).
6. M. P. Schutzenberger, "On Context-free Languages and Pushdown Automata," *IBM Res. Rept. RC-793* (1962); A. G. Oettinger, "Automatic syntactic analysis and the pushdown store," *Proc. Symp. Appl. Math.* **12**, 104 (1961).
7. J. Eickel, M. Paul, F. L. Bauer, K. Samelson, "A syntax-controlled generator of formal language processors," *Commun. Assoc. Computing Machinery* **6**, 451 (1963).
8. J. P. Anderson, "A computer for the direct execution of algorithmic languages (B-5000)," *Proc. Eastern Joint Computer Conf.* **20** (1961); A. C. D. Haley, "The KDF-9 computer system," *Proc. Fall Joint Computer Conf.* **22**, 108 (1962).
9. P. Dreyfus, "Programming design features of the Gamma-60 computer," *Proc. Eastern Joint Computer Conf.* **14** (1958).
10. T. Kilburn, D. B. G. Edwards, M. J. Lanigan, F. H. Sumner, "One-level storage systems," *Trans. Electronic Computers, Inst. Elec. Electronic Engrs.* **EC11**, No. 2, 223

(1962); T. Kilburn, R. B. Payne, D. J. Howarth, "The Atlas Supervisor," *Proc. Eastern Joint Computer Conf.* **20** (1961); D. J. Howarth, "Experience with the Atlas scheduling system," *Proc. Spring Joint Computer Conf.* **23** (1963); ———, P. D. Jones, M. T. Wyld, "The Atlas scheduling system," *Computer J.* **5**, 238 (1962).

11. D. L. Slotnick, W. C. Borck, R. C. McReynolds, "The Solomon computer," *Proc. Fall Joint Computer Conf.* **22**, 97 (1962); J. K. Hawkins and C. J. Munsey, "A parallel computer organization and mechanizations," *Trans. Electronic Computers, Inst. Elec. Electronic Engrs.* **EC12**, 251 (1963); C. Y. Lee and M. C. Paull, "A content addressed distributed logic memory with applications to information retrieval," *Proc. Inst. Elec. Electronic Engrs.* **51**, 924 (1963).
12. J. W. Lewis, "Time sharing on Leo III," *Computer J.* **6**, No. 1, 24 (1963); M. J. Marcotty, F. M. Longstaff, A. P. M. Williams, "Time sharing on the Ferranti Packard FP6000 computer system," *Proc. Spring Joint Computer Conf.* **23** (1963).
13. C. Green, "The Euratom Computer Linkage System," *Euratom Rept. EUR 284e* (1963); A. Debroux, G. P. del Bigio, A. Gazzano, C. Green, H. d'Hoop, A. Riotte, A. Van Wauwe, "Utilization of an analogue-to-digital linkage system in a big scientific computing center," in *Proc. Intern. Fed. Inform. Processing Congr., 1962* (North-Holland, Amsterdam, 1963), pp. 236–241.
14. N. Chomsky and M. P. Schutzenberger, "An algebraic theory of context-free languages," in *Computer Programming and Formal Systems*, P. Braffort and D. Hirschberg, Eds. (North-Holland, Amsterdam, 1963), pp. 118–161; H. W. Gumin, "The influence of programming languages on the organization of digital computers," in *Proc. Intern. Fed. Inform. Processing Congr., 1962* (North-Holland, Amsterdam, 1963), pp. 566–569.
15. P. Lucas, "Die Strukturanalyse von Formelübersetzern," *Electron. Rechenanlagen* **3**, 159 (1961); R. A. Brooker and D. Morris, "An assembly program for a phrase-structure language," *Computer J.* **3**, 168 (1960); ———, "A general translation program for phrase-structure languages," *J. Assoc. Computing Machinery* **9**, 1 (1962); R. A. Brooker, I. R. MacCallum, D. Morris, J. S. Rohl, "The compiler compiler," in *Annual Review in Automatic Programming, No. 3*, R. Goodman, Ed. (Pergamon, London, 1963), pp. 229–276; E. N. Hawkins and D. H. R. Huxtable, "A multi-pass translation system for Algol-60," *ibid.*, pp. 163–206; M. Paul, "Algol-60 processors and a processor generator," in *Proc. Intern. Fed. Inform. Processing Congr., 1962* (North-Holland, Amsterdam, 1963), pp. 493–497.
16. D. W. Barron, J. N. Buxton, D. F. Hartley, E. Nixon, C. Strachey, "The main features of CPL," *Computer J.* **6**, 134 (1963); E. W. Dijkstra, "On the design of programming languages," in *Annual Review in Automatic Programming, No. 3*, R. Goodman, Ed. (Pergamon, London, 1963), pp. 27—42.
17. V. Kudielka, P. Lucas, K. Walk, K. Bandat, H. Bekic, H. Zemanek, "Extension of the algorithmic language Algol," *Final Rept. DA-91-591-EUR-1430, IBM Science Group, Vienna* (1961).
18. P. Lucas, "Requirements on a language for logical data processing," in *Proc. Intern. Fed. Inform. Processing Congr., 1962* (North-Holland, Amsterdam, 1963), pp. 556–559.
19. R. Tabory, "Premiers elements d'un langage de programmation pour le traitement en ordinateur de graphes," in *Symbolic Language in Data Processing* (Gordon and Breach, London, 1962), pp. 717–730.
20. K. H. Meyer-Uhlenried and C. Muyzenberg, "Use of macro-instructions and program generators in automatic documentation," in *Proc. ADI Annual Meeting, 1963* (American Documentation Institute, Chicago, 1963), p. 29.
21. H. P. Schlaeppi, "A formal language for describing machine logic, timing, and sequencing (LOTIS)" (I.B.M. Research Laboratory, Zurich, 1963).
22. G. Veillon, "Consultation d'un dictionnaire et analyse morphologique en traduction automatique," thesis, University of Grenoble (1962).
23. H. Schnelle, "Programmieren Linguistischer Automaten," in *Neuere Ergebnisse der Kybernetik*, K. Steinbuch, Ed. (Oldenbourg, Munich, 1963).
24. B. Vauquois and J. Veyrunes, "Presentation de l'Analyse Morphologique du Russe," University of Grenoble Rept. G-100-C (University of Grenoble, 1962); G. Veillon, "Presentation de l'Analyse Morphologique et du Programme de Dictionnaire Allemand," University of Grenoble Rept. G-500-A (University of Grenoble, 1963); B. Vauquois, "Langages Artificiels—Systèmes Formels et Traduction Automatique," course presented at the NATO Advanced Study Institute, Venice (1962).
25. J. Blois, F. Decresy, J. Mommens, "Analyse Morphologique Automatique du Français," report to Euratom under contract 018-61-CET-B, University of Brussels (1963).
26. J. A. Bachrach, J. Blois, F. Decresy, F. Defijn, L. Hirschberg, J. Mommens, "Dicautom-Consultation Automatique de Dictionnaires pour Traducteurs Humains," report to Euratom under contract 018-61-5-CET-B, University of Brussels (1963).
27. L. Tesnière, *Eléments de Syntaxe Structurale* (Klincksieck, Paris, 1959).
28. L. Hirschberg and I. Lynch, "Discussions sur l'Hypothèse de Projectivité," *CETIS Rept. No. 35, Euratom, Ispra* (1961); Y. Lecerf, "Programme des Conflits—Modèle des Conflits," *CETIS Rept. No. 4, Euratom, Ispra* (1960); E. Scheffer, "Recueil de Stemmas," *CETIS Rept. No. 29, Euratom, Ispra* (1961).
29. L. Hirschberg, "Ponctuations et Analyse Syntaxique Automatique," report to Euratom under contract 018-61-5-CET-B, University of Brussels (1962); "Punctuation and Automatic Syntactic Analysis," paper presented at annual meeting of the Association for Machine Translation and Computational Linguistics, Denver (1963).
30. S. Ceccato *et al.*, "Mechanical Translation: "The Correlational Solution," technical report to European Office of Air Force Office of Scientific Research by Centro di Cibernetica, Milan (1963); E. Glasersfeld, S. Perschke, E. Morpurgo, "Travaux du Centro di Cibernetica et di Attivita Linguistiche," *CETIS Rept. No. 24, Euratom, Ispra* (1961).
31. S. Ceccato *et al.*, "An example of mechanical translation," paper presented at the Convegno Europeo per Redattori Scientifici, Rome (1962).
32. R. M. Needham, "A method for using computers in information classification," in *Proc. Intern. Fed. Inform. Processing Congr., 1962* (North-Holland, Amsterdam, 1963), pp. 284–287; A. F. Parker-Rhodes and R. M. Needham, "The Theory of Clumps," *Cambridge Language Research Unit, Cambridge, Repts. ML138, ML139* (1961).
33. R. M. Needham, "Automatic classification of index terms and documents," paper presented at the NATO Advanced Study Institute for Automatic Document Analysis, Venice (1963).
34. A. R. Meetham, "Preliminary studies for machine-generated index vocabularies," *Language and Speech* **6**, pt. 1, 22 (1963).
35. F. Levery, "Une Experience d'Indexage Automatique," *IBM France Rept.* (1963).
36. H. Buntrock and K. H. Meyer-Uhlenried, "Terminology work for mechanized and semi-mechanized documentation systems," in *Proc. ADI Ann. Meeting, 1963* (American Documentation Institute, Chicago, 1963), pp. 19–20; K. H. Meyer-Uhlenried and G. Lustig, "Analysis, indexing, and correlation of information," *ibid.*, pp. 229–230.
37. L. Rolling, "Un Repertoire de Mots-clés pour la Documentation dans le Domaine de la Technique Nucleaire," *Euratom Rept. EUR 227.f* (1963); E. Pietsch, AED—Atomic Energy Information Service, Gmelin Institute, Frankfurt (1963).
38. J. Ruvinschii, "Consignes Provisoires pour la Mise en Diagrammes des Textes Scientifiques," *CETIS Rept. No. 5, Euratom, Ispra* (1960); Y. Lecerf and A. Leroy, "Description d'un Algorithme d'Analyse Documentaire, *CETIS Rept. No. 6, Euratom, Ispra* (1960); Y. Lecerf, "Programmes de Conflits," *Euratom Rept.* (1960).
39. J. C. Gardin and F. Levy, "Le Syntol-Syntagmatic Organization Language," in *Proc. Intern. Fed. Inform. Processing Congr., 1962* (North-Holland, Amsterdam, 1963), pp. 279–283; J. C. Gardin *et al.*, "A General System for the Treatment of Documentary Data," *Final Rept. to Euratom, Assoc. Marc Bloch, Paris* (1962); M. Coyaud, "Analyse automatique de documents ecrits en langue naturelle vers un langage documentaire (Syntol)," paper presented at the NATO Advanced Study Institute for Automatic Document Analysis, Venice (1963).
40. R. Busa, "The use of punched cards in linguistic analysis," in *Punched Cards—Their Application to Science and Industry*, R. S. Casey *et al.*, Eds. (Reinhold, New York, 1958), pp. 357–373; B. Quemada, *Actes du Colloque International sur la Méchanisation des Recherches Lexicologiques* (Didier-Larousse, Paris, 1962); H. Schnelle, "Uber den Stand der Forschung zur automatischen Sprachbearbeitung im deutschen Sprachraum," in *Sprachkunde und Informationsverarbeitung* (Oldenbourg, Munich, 1963), vol. 2, pp. 48–61.
41. J. Iung *et al.*, "Bibliographie Auto-indexée—Physique des Plasmas et Fusion Thermonucléaire Controlée," *Publ. Service Central de Documentation du CEA* (1962); N. Chonez, A. Chonez, J. Iung, "Physindex: An auto-indexed current list of physics literature produced on IBM 1401 computer," in *Proc. ADI Ann. Meeting, 1962* (American Documentation Institute, Chicago, 1963); N. Vandeputte, "Traitement sur IBM 1401 de Textes Scientifiques Anglais en vue d'Etudes Linguistiques et Statistiques," *CEA Rept. No. 369, Service de Documentation du CEA* (1961).
42. F. Wegmüller, *Literaturübersicht—Codeless Scanning* (Hoffman La Roche, Basle, 1963).
43. ———, R. Becker, B. Hoffman, H. R. Schenk, "Codeless scanning, Ein Neues Verfahren der Automatischen Dokumentation," in *Separatum Experientia* **16**, No. 8, 383 (Birkhauser, Basle, 1960).
44. T. W. te Nuyl, "The l'Unité Mechanized Documentation System," *Rev. Intern. Doc.* **28**, No. 4, 140 (1961); "The l'Unité Documentation System," *ibid.* **25**, No. 3, 65 (1958).
45. E. Meyer, "Encoding of organic-chemical structural formulas and reactions by machine," in *Proc. Ann. ADI Conv. 1963* (American Documentation Institute, Chicago, 1963), pp. 131–132.
46. J. Gevers, *Les Marques de Fabrique et de Commerce* (Gevers, Antwerp, 1963).
47. C. W. Cleverdon and J. Mills, "The testing of index language devices," *Aslib Proc.* **15**, 106 (1963); J. Aitchison and C. W. Cleverdon, "A Report on a Test of the Index of Metallurgical Literature of Western Reserve University," *Aslib-Cranfield Project Rept.* (1963); C. W. Cleverdon and J. Mills, "The analysis of index language devices," in *Proc. ADI Ann. Conv. 1963* (American Documentation Institute, Chicago, 1963).