

[from: *Discourse and Communication: new approaches to the analysis of mass media discourse and communication*, ed. Teun A. van Dijk (Berlin: Walter de Gruyter, 1985)]

W. JOHN HUTCHINS

## Information Retrieval and Text Analysis

### *1. Introduction*

All scientists and scholars know that the effective dissemination and reception of new ideas, theories and experimental results and the open dispassionate discussion and evaluation of hypotheses and research findings are absolutely crucial factors in the advancement and intellectual vitality of any field of scientific or scholarly activity. Access to what their fellow scientists and scholars have said and written is essential not only for individual researchers but also for the evolution of that consensus of opinion which represents the corpus of 'knowledge' in the discipline (Ziman, 1968; Popper, 1972). In this communication network the published literature plays a central role. It is virtually a truism that 'scientific knowledge' exists primarily in the documentation of science (journal articles, reports, conference proceedings, dissertations, textbooks, etc.) and only secondarily in the fallible memories of individual scientists. It is also true to say that scientists and scholars exist professionally (i. e. as researchers and thinkers, rather than as teachers or administrators) by and through their contributions to the literature of their subjects and by the influence of their publications on other scientists and scholars. The maintenance of this system requires and has always required effective ways and means of gaining access to and finding out about what has been published, i. e. effective 'information retrieval'.

In essence, information retrieval means the extraction of 'information' of some kind (data, texts, references) from a 'store' (memory, file, database, document collection) in response to an 'information need'. The term is now often used for systems which do not involve published or recorded information, such as biological and psychological processes in animal and human memories and computer models of 'human information processing'. This paper is concerned with the more traditional concept of information retrieval, with the problems of handling the written and recorded texts and data of scientific documentation. A general overview of the processes involved is given in figures 1 and 2.

Systems may respond in basically two ways to requests: either by providing the actual data which satisfies an information need, or by providing citations of documents (i. e. details of authors, titles, journals, etc.) which might contain

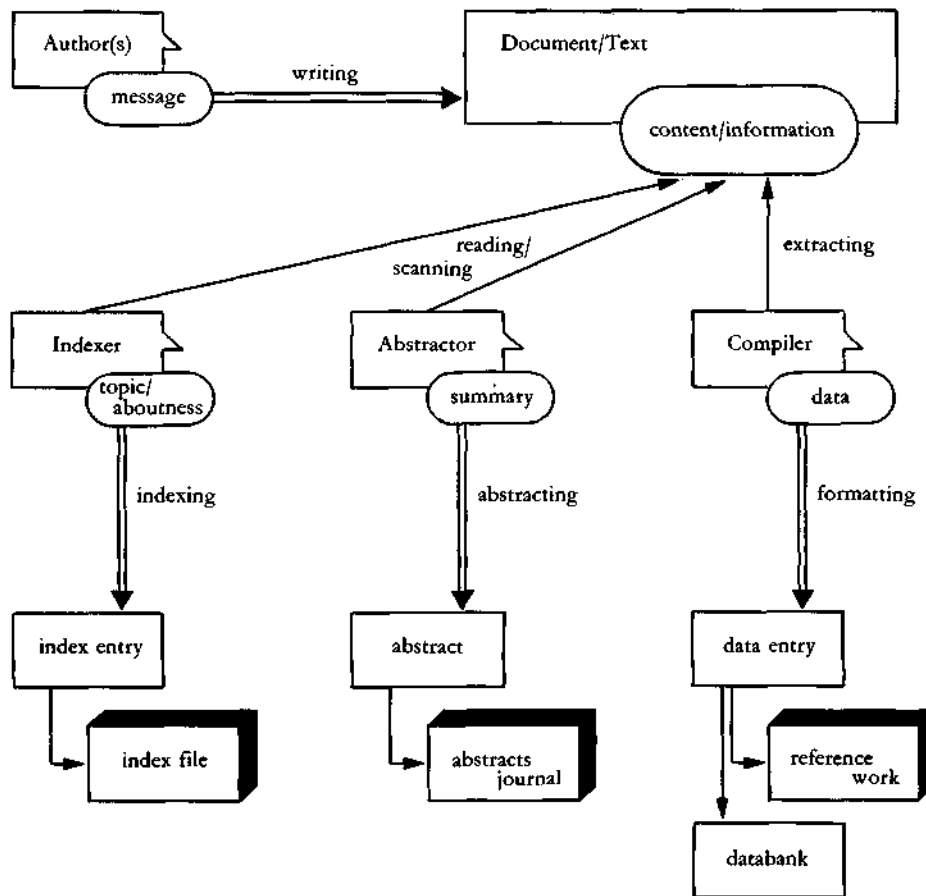


Fig. 1: Information storage

the information sought. The first are commonly referred to as 'fact-retrieval' or 'data-retrieval' systems if they are computer-based; their more familiar printed equivalents are dictionaries, directories, lists of addresses, handbooks of mathematical tables, physical constants, etc. The latter are commonly referred to as 'bibliographic' or 'document-reference' systems; their familiar forms are catalogues, indexes, bibliographies and abstracts journals. Information about the subject contents of documents may be recorded in bibliographic systems either in 'index entries' (individual words or phrases or 'terms' stating the topics treated or mentioned in documents), or in 'abstracts' (brief paragraphs summarizing the essential 'messages' of documents), or in extracts from actual texts (most commonly titles and subtitles, but sometimes longer passages).

Figure 1 represents the processes involved in the production of records or data for information retrieval systems: the writing of texts by authors to express a 'message'; the indexing of texts by indexers to express what they are 'about'; the abstracting of texts by abstractors to express the essence of authors' messages; the extracting of data from texts (and other sources) by compilers of reference works and databases. Figure 2 represents the processes involved in the searching of records and data: the expression of information needs; the formulation of search requests appropriate to the type of record (index, abstract, database); the extraction of information from the system; the evaluation of the information retrieved.

Text analysis and text production are clearly at the heart of information retrieval systems: indexers and abstractors must read and understand the texts they are indexing and abstracting; compilers of dictionaries, handbooks and databases must analyze and arrange ('format') the information to be included; abstractors must write the texts of abstracts; users of systems must express their needs in texts (search requests), and they must interpret the texts retrieved for them. Much of this text processing is not at all peculiar to

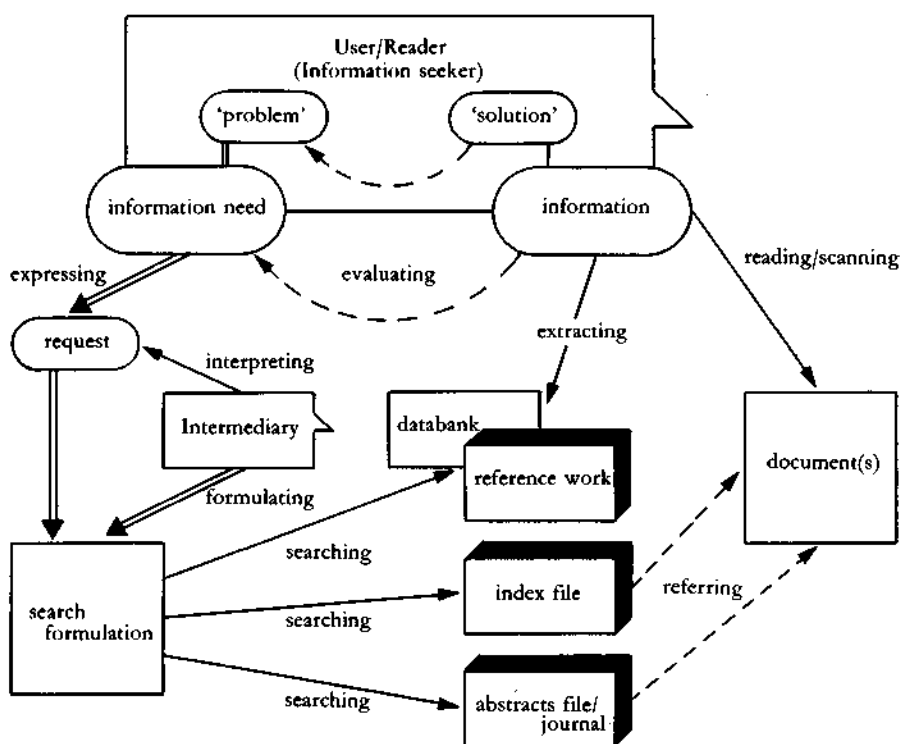


Fig. 2: Information searching

information systems: the ways that indexers, abstractors and information seekers read and interpret documents are no different from those of readers in general; and the ways that abstractors write abstracts and that users express their search requests are no different in essence from those of authors and speakers in general. What is peculiar to information retrieval systems are the processes of summarizing, condensing and extracting texts and the processes of matching search requests against texts (titles, abstracts) or text-like representations (index entries, structured databases) — these are the principal topics of this paper.

An attempt is made to present a unified picture by relating what is known or surmised about human indexing, abstracting and searching to accounts of computer-based systems of automatic indexing, abstracting and searching. Section 2 concentrates on the problems of natural language access to systems, sections 3 and 4 on the analysis of texts to produce indexes and abstracts, section 5 on the 'formatting' of information in data-retrieval systems, and section 6 on the problems of evaluating the results of searches.

## 2. Searching

Whether searching 'bibliographic' or 'data-retrieval' systems, users are required to express their information needs as precisely as possible in the form appropriate to the system concerned. It is a difficult task and one which is little understood. There may be four stages (Taylor, 1968): an initial awareness of a 'problem space' (a lack of understanding, a dissatisfaction with current explanations, an anomaly in the data, etc.); a preliminary, informal, partly incoherent expression of a 'need'; a more coherent, rational and considered statement of the kind of information or document which might satisfy the need; and lastly, a search request formulated in the 'language' of the information system.

2.1. In the case of non-computerized systems, users generally require the assistance of 'intermediary experts' (e. g. librarians, information officers) to adapt requests and to search systems. Searching indexes, for example, demands knowledge of the way index entries are constructed and how they are to be interpreted if users are to make the most effective use of them. (Index languages have indeed their own semantic and syntactic structures and may be studied as linguistic systems in their own right (Gardin, 1973; Hutchins, 1975).) Searching 'formatted' sources makes equivalent demands on users; they have to understand the structure and meanings of tables (e. g. of statistical data), the norms and conventions of dictionary entries, the formats in which biographical information is presented, etc.

The process of adapting (even 'distorting') requests to the particular requirements of individual information systems is undoubtedly one of the

reasons why all too often users fail to find the data or documents that satisfy their needs. What is not always realised is that the very process of producing a coherent statement of 'information need' contributes a certain degree of distortion. By its very nature an information need cannot be expressed precisely and concisely. It may well be plausible to suggest (Beaugrande and Dressler, 1981) that in 'normal' text production speakers and writers have some idea of what they want to say or write, and adopt suitable plans and strategies for putting over their 'messages' — even if they have often to make corrections and revisions when they realise that they have not said or written what they intended. In the expression of information needs, however, the problem is that users do not know precisely what they want at any stage — there is no clear 'message' or 'intention' to go back to, only a vague awareness of a 'problem area'. Requests are inevitably imprecise and approximate, and nobody knows whether they correspond to 'real' information needs or not — hence, in part, the difficulties of evaluating searches (cf. section 6).

2.2. Computerized bibliographic systems enable users to search not only the index terms assigned to documents but also natural language texts: the titles and abstracts of documents in nearly all cases, but in some systems even the full texts of documents (e. g. the legal retrieval systems LEXIS and FLITE; cf. Larson, 1977) and as computer storage capacities increase and more and more documents and records become available in machine-readable form (often as by-products of printing/publishing processes) this is likely to be the norm in future systems (Lancaster, 1977). Yet paradoxically, the users of these operational systems cannot formulate requests in 'free' natural language; they must break down requests into individual terms, for which they must provide possible synonyms (or near-synonyms) and alternative spellings, and decide how terms are to be combined (present systems allow only Boolean links: *and, or, not*). Assistance may be provided by manuals and thesauri and also by 'intermediary experts', but what these bibliographic systems clearly need are more 'friendly' interfaces. Considerable research has been devoted to systems which help users to modify and refine searches — mainly by statistical techniques which adjust search formulations (e. g. adding or subtracting terms, substituting more general or more specific terms, etc.) in the light of users' evaluations of a selection of retrieved documents (cf. Rijsbergen, 1979; Bates, 1981). There is now also increasing research activity on systems which accept requests in 'free' natural language, which identify search terms and supply synonyms and alternatives and which enable users to interact with the system during searches. One of the most advanced systems of this kind is CITE (Doszkocs and Rapp, 1979), designed for searching the large medical database MEDLINE; this incorporates not only automatic identification of search terms and automatic construction of search formulae but also interactive modification during searches and ranking of documents according to potential 'relevance' in meeting users' needs.

However, efficient natural language interfaces will not solve all difficulties in searching bibliographic databases. For example, there is the inevitable 'distortion' of users' information needs by the inherent constraints of information systems. Belkin and Oddy (1979) suggest that searching should be based not on (relatively precise) search requests, however 'freely' expressed, but on informal, rambling, loose (even incoherent) statements by users of their 'problem' as they see it and how they think it might be 'solved' (i. e. Taylor's second stage in the expression of 'information need'); such statements would provide an outline 'profile' of search terms to be modified and made more precise during interactive searching. Another problem is the imprecision of bibliographic searches (i. e. too many irrelevant documents retrieved) which may result from failure to take into account semantic and syntactic relationships within document texts. The problem becomes acute if the full texts of documents are searched, since clearly search terms (i. e. individual words) may be found within a document in quite unrelated passages. One approach is illustrated by the research of O'Connor (1975, 1980) on full text searching. Documents are retrieved only if they contain passages of text which both include a minimum number of search terms and consist of 'connected' sentences. Connectivity is determined by the occurrence of connector words (e. g. conjunctions and sentence adverbs) and by the repetition of search terms. For example, in response to 'What are the effects on the ureter of radiation treatment of cancer?' the following passage was located (words corresponding to search terms are italicized):

Complications of *radiation therapy* for prostatic *carcinoma* are frequent but usually minor. A patient is described in whom localized *ureteral* fibrosis developed following *curative radiation therapy*, which subsequently required a reconstructive operation. (O'Connor, 1980: 233)

2.3. Most research on natural language interfaces for database searching has taken place in the context of artificial intelligence work on question-answering systems (i. e. fact-retrieval systems). One of the best known examples is LUNAR, developed by Woods (1977, 1978) to enable geologists at NASA to interrogate a structured ('formatted') database containing results of chemical analyses of lunar rock and soil samples from the Apollo missions. A request such as 'How many samples contain silicon?' is first parsed (by the now familiar Augmented Transition Network parser):

S	Q		
	NP	DET	HOWMANY
		N	SAMPLE
		NU	PL
	AUX	TNS	PRESENT
	VP	V	CONTAIN
		NP	DET NIL
		N	SILICON
		NU	SG

and then converted into a logical format specifying the procedures (e.g. calculations) to be carried out on data from the file. Thus, HOWMANY is interpreted as an instruction to calculate the number of records satisfying the specified conditions (presence of silicon), and to print out the result. However, LUNAR has very limited facilities for user-system interaction. Later systems enable searchers to conduct more or less natural language dialogues — within the subject domain of the database. For example, GUS (Bobrow, et al., 1977) acts as a travel agent booking airflights in California:

CLIENT: I want to go to San Diego on May 28  
 GUS: What time do you want to leave?  
 CLIENT: I must be in San Diego before 10 am  
 GUS: Would you like P.S.A. flight 102 that arrives at 9.15 am?  
 CLIENT: What is the next flight?  
 GUS: Air California flight 310 that leaves at 8.30 and arrives at 10.00 am  
 CLIENT: I'll take the first one

The example illustrates the handling of complex problems of anaphora: *next flight* refers both to GUS's preceding statement and to the order of flights in the airline guide, and *first one* refers to the order in which flights have been mentioned during this particular dialogue. A number of question-answering systems are able to handle ungrammatical and misspelled requests and to deal with elliptical expressions (e.g. Waltz, 1978) — facilities which are now regarded as essential for 'intelligent' interface systems (Weischedel and Black, 1980).

One system incorporating all these features is the LIFER natural language interface as applied in a highly complex database network providing information for the U.S. Navy (Hendrix, et al., 1978). An illustrative interchange (with comments in parentheses) is given in figure 3. Like most question-answering parsers LIFER embeds considerable semantic information in its syntax — it is a 'semantic grammar'. For example, words like NAUTILUS and DISPLACEMENT are not grouped together in a single Noun category, but treated as /SHIP-NAME/ and as /ATTRIBUTE/ respectively; and, instead of general syntactic patterns such as *NP + VP*, it analyzes in terms such as:

WHAT IS THE /ATTRIBUTE/ OF /SHIP-NAME/

As a result, 'semantic grammars' tend to be domain-specific and not easily transferred to other databases. However, LIFER is more flexible than most — it has been applied also as an interface to a medical database (Epstein and Walker, 1978) — as it has been developed within the context of a large-scale programme of information retrieval research (Walker, 1981) in order to provide natural language access to a variety of different kinds of databases. Increased flexibility and domain-independence is the goal of many current projects; it has been achieved to some extent by Sager's linguistic string parser (section 5.2 below) and by the ROBOT parser (Harris, 1977). Already applied to six different databases of great variety, ROBOT derives its semantic

1. What is the length of the Constellation  
 PARSED! (= request successfully analyzed,  
 (LENGTH 1072 feet) database being searched)
2. of the Nautilus  
 TRYING ELLIPSIS: WHAT IS THE LENGTH OF THE NAUTILUS  
 (LENGTH 319 feet) (= expanded request successfully answered)
3. displacement  
 TRYING ELLIPSIS: WHAT IS THE DISPLACEMENT OF THE NAUTILUS  
 (STANDARD-DISPLACEMENT 4040 tons)
4. length of the fastest American Nuclear sub  
 TRYING ELLIPSIS: WHAT IS THE LENGTH OF THE FASTEST AMERICAN  
 NUCLEAR SUB  
 (LENGTH 360 feet NAME LOS ANGELES SPEED 30.0 knots)
5. Who commands the Constellation  
 SPELLING -- CONSTELLATION (= correction of misspelling)  
 PARSED! (= corrected form analyzed)  
 (COMMANDER CAPT J. ELLISON)
6. Who commands JFK  
 TRYING ELLIPSIS: ELLIPSIS HAS FAILED  
 THE PARSER DOES NOT EXPECT THE WORD "JFK" TO FOLLOW "WHO  
 COMMANDS"  
 OPTIONS FOR NEXT WORD OR META-SYMBOL ARE: /SHIP-NAME/
7. Define JFK to be like Kennedy  
 PARSED! (= acknowledges user's definition of 'JFK'  
 as a synonym of KENNEDY, which is  
 already known as a /SHIP-NAME/.)
8. Who commands JFK  
 PARSED! (= retry of 6 is now accepted)  
 (COMMANDER CAPT P. MOFFETT)

Fig. 3: LIFER dialogue

independence largely from utilizing dynamically the structural properties of the database itself during analysis and interpretation. The advent of more flexible and independent interfaces and their application to larger databases promises improved natural language access not only to fact-retrieval systems but also to bibliographic databases.

### 3. Indexing

In bibliographic systems the index terms associated with documents may either be 'derived' from words or phrases occurring in the documents, i. e. extracted from actual texts, or they may be 'assigned' (selected) by indexers or indexing systems. The selection may be 'controlled' in that only terms appearing in a

particular list (an authority file) may be used, or it may be 'uncontrolled'. Index terms for a particular document may, in some systems, refer to the 'topic' of the document considered as a whole ('topic indexing'), or they may, in other systems, refer to any subjects the document may mention so that, taken together, they constitute a kind of 'summary' of the document's message ('summary indexing'). In addition, indexing may be 'user-oriented' in that only those terms are assigned which are expected to be sought for by the particular clients of the system; or it may be 'document-oriented' if general, 'objectively' valid characterizations are intended which are not specific to particular environments.

3.1. The indexing done by human indexers is invariably 'assigned' indexing, most often with 'controlled' terms. In systems for the general user (public libraries, general bibliographies) the usual approach is that of 'topic indexing' with a predominantly 'document-orientation'. In systems for specialists and experts (research libraries, specialised bibliographies) the usual approach is that of 'summary indexing' with a 'user-orientation'.

Little is known about how indexing is done; guides and manuals for indexers concentrate on formal properties of index terms and the construction of index entries; they say nothing about how indexers decide what documents are 'about' or how they select suitable index descriptions. In 'user-oriented' indexing it is suggested that they scan texts for particular words or phrases known to be likely search terms in the relevant specialism, i. e. they refer to an internalised check list (Soergel, 1974: 47). As for 'topic indexing' it is suggested that the theme-rheme articulations of paragraph and text structures (cf. Daneš, 1974) provide clues to global topics of documents and that these may be related to the 'given' knowledge which authors assume their potential readers have already (Hutchins, 1978). By contrast, in 'summary indexing' the notion of topic appears to be related to the node in a text-linguistic representation of a paragraph (or text) which has most links to other nodes, i. e. generally a noun phrase occurring as subject in more than one sentence and referred to by pro-forms — e. g. *rocket* in the following passage analyzed by Beaugrande and Dressler (1981: 103):

With a great roar and burst of flame the giant rocket rose slowly at first and then faster and faster. Behind it trailed sixty feet of yellow flame. Soon the flame looked like a yellow star. In a few seconds, it was too high to be seen, but radar tracked it as it sped upward to 3,000 mph.

3.2 Most experimental work on automatic indexing has concentrated on statistical methods of analysis (Sparck Jones, 1974; Harter, 1978), the general assumption being that frequency of occurrence is correlated broadly with semantic importance. Since the crude counting of word tokens is clearly inadequate (in the passage above, *flame* occurring three times would be ranked

higher than *rocket* occurring once), many subtle and complex refinements are employed: e. g. the exclusion of words which occur frequently in the subject field (as well as words frequent in the language in general, such as function words), the truncation of words to bring together morphologically related words (suffix-stripping), the normalisation of frequencies to allow for varying document lengths, the weighting of words appearing in certain 'important' parts of texts (titles, section headings, conclusions), the use of co-occurrence frequencies, and so forth.

In general the terms derived by automatic indexing are 'uncontrolled', but there are exceptions. Klingbiel (1973) and Barnes et al. (1978) describe systems where potential terms are scrutinized against an authority file: some are accepted, some rejected, some replaced by synonyms, others replaced by more general terms. Another form of control is described by Sparck Jones (1971): if two terms tend to co-occur in the same documents then they may be equally effective as search terms for the retrieval of those documents, i. e. they are mutually substitutable (whether or not they happen to be close in meaning and whether or not they happen to refer to similar topics). A good discussion of statistical methods in information retrieval is to be found in Rijsbergen (1979).

3.3. Research in automatic indexing has made relatively little use of linguistic methods of analysis. Some limited parsing is found in the systems of Klingbiel (1973) and Barnes et al. (1978), where a simple parser identifies nouns and noun phrases for testing as potential index terms, and in the LEADERMART system (Hillmann, 1968, 1973), where a parser identifies nouns and 'logical' relations between them (from analyses of prepositions, conjunctions and simple phrase structures). However, two projects have been more ambitious: the SMART and SYNTOL systems of automatic indexing.

SMART now uses statistical methods almost exclusively, but in earlier versions (Salton, 1968) there was some semantic and syntactic analysis. Each sentence of a document (or, more often, its abstract) was parsed by the Harvard Predictive Analyzer (a finite-state parser designed originally in the 1950's for machine translation), which produced a basic phrase structure analysis in dependency grammar format. From the parsing were extracted substructures (e. g. subject-verb and noun-adjective links) to be matched against a dictionary of 'criterion phrases'. A criterion phrase was a dependency tree in which each node represented a set of semantically related words (a concept group) and each link a defined syntactic relationship (e. g. attribute, instrument), and which could be treated in subsequent procedures as a single unit. There were considerable problems in establishing the syntactic and semantic conditions for criterion phrases, but there were even more difficulties with the inadequacies of the parser, which notoriously produced either no analyses at all or far too many — and it is not surprising that Salton abandoned syntactic analysis in favour of the then more satisfactory statistical methods.

In certain respects, the analytical procedures of SYNTOL (Bely, et al., 1970) were similar: here too, the parser (context-free) produced a dependency structure from which substructures could be extracted representing pairs of index terms linked by basic relations. These were the 'syntagms' which were combined in networks in order to represent the content of document texts (or rather, document abstracts). It was realised that if syntagms were to be found successfully in searches they would not have to be too specifically defined; in fact they were perhaps made too abstract (eventually only three types of link were permitted) and the analysis program was not powerful enough to convert the semantic complexities of the natural language input into the required abstractness of SYNTOL's syntagmatic representations.

#### *4. Abstracting*

Research on automatic abstracting has tended to be more ambitious from a linguistic point of view than most research on automatic indexing. The reason is not far to seek: not only must the texts of documents be analysed in some detail, but also texts must be produced (the abstracts) which are coherent syntactically and semantically and which at the same time are reasonable 'summaries' of some kind of the original documents.

*4.1.* The process of 'summarization' is itself highly complex, as we shall see, but it is clear from manuals and guides for abstractors (e. g. Bernier, 1968; Weil, 1970) that abstracts are more than just summaries. Other text types also include condensations of texts, e. g. review articles surveying the literature of a subject, newspaper articles reporting research, handbooks outlining the current 'state of knowledge' in a discipline, encyclopaedia articles, etc. (Bernier, 1970); and abstracts are distinct from these. Two basic types are identified: 'informative' abstracts which include actual results, figures and conclusions from source documents, and 'indicative' abstracts which simply record the fact that certain topics are covered. (The distinction is roughly parallel to that between 'summary indexing' and 'topic indexing'.) To maintain standards and consistency, recommendations and guidelines for abstractors are often very specific: state the purpose of the work reported; give the methods used, the results obtained and the conclusions reached; retain the balance and emphases of the original; convey only what is 'new' information; pass no comments, either favourable or critical; produce a self-contained coherent text within the confines of (ideally) a single paragraph and which might stand as a substitute for the original for some purposes; and so forth. The linguistic complexities of abstracting clearly transcend just 'summarizing' and it is not surprising that they still await investigation.

Summarization itself is complex enough as a linguistic process. Within his general theory of text linguistics, Van Dijk (1977, 1980) has outlined some

basic operations. Van Dijk distinguishes between the microstructure of a text (the underlying propositional content of its sentences and clauses, and their connections to each other, in the linear sequence in which they are expressed) and its macrostructure (the semantic representation of the text as an entity, independently of its particular propositional manifestation). Summaries are one way of expressing the macrostructures of texts. Macrostructures are derived from microstructures by the operations of four types of 'macro-rules'. Two are concerned essentially with the identification of 'important' propositions: *deletion* operates negatively by eliminating the unnecessary and irrelevant (e.g. detailed descriptions, background information, common knowledge), and *selection* operates positively by extracting the necessary and the relevant (e.g. propositions expressing pre-conditions and data required for the interpretation of other propositions). The other two are concerned with condensation and abstraction: *generalization* constructs general propositions from the semantic detail of microstructural propositions (e.g. from a description of girls playing with dolls, boys playing with train sets, etc. it derives a description of children playing with toys), and *construction* replaces sequences of propositions by single propositions expressing self-contained events or processes ('scripts').

4.2. The experience of research on automatic abstracting indicates that deletion and selection are more easily simulated in computer analysis than are the other summarization operations. Initially, researchers attempted to produce 'abstracts' by extracting sentences en bloc from texts on the basis of high frequency words (excluding function words and items of common vocabulary), e.g. Luhn's (1958) pioneering work. The results were neither particularly good condensations nor very coherent texts. Later systems have combined more sophisticated statistical methods (similar to those in automatic indexing) with attempts to use the kind of textual 'cues' which abstractors seem to use. Edmundson (1969) and Rush et al. (1971) employed three types of 'cues': (i) the recurrence of words in titles, subtitles and section headings, or the occurrence of words synonymous with them, (ii) the presence of words such as *significant*, *impossible*, *hardly*, which indicate authors' views of the importance of the information presented, and (iii) the location of sentences within paragraphs and sections. Such 'cues' identify potentially 'important' passages — they are in part refinements of Baxendale's (1958) observations on the occurrence of topic sentences in paragraphs and the overt textual marking of important passages. (Similar observations are part of the rhetorical tradition now formalised to some extent in text linguistics.)

Most researchers concede that these procedures should be properly called 'automatic extracting' of sentences; but there have been some attempts to produce coherent sequences, to make some generalizations and so to come closer to 'abstracting' of some kind. Mathis et al. (1973) introduced various refinements into the procedures of Rush et al. (1971): (i) if an extracted

sentence contains an anaphoric link to a preceding sentence then the latter is also included in the abstract, (ii) specific references to particular tables, graphs, etc. (e. g. *Table 2, figure 3, the second mechanism*) are changed to general references (*a table, a figure, a mechanism*), and (iii) some extracted sentences are combined by coordinate and subordinate conjunctions. The latter involved limited parsing of candidate sentences to identify noun, verb, and preposition phrases, to locate antecedents of pronouns, and to recognize parallel structures. For example, the two sentences:

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

could be combined as:

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

Obviously, much more is needed than such simple syntactic manipulations if genuine abstracting and summarizing is to be achieved; some indications of the techniques required are to be seen in artificial intelligence, where the concept of 'script' in text understanding appears to be an example of Van Dijk's macro-rule of construction.

4.3. Most pertinent in this context is the research of Schank and his colleagues (Schank, 1975) on programs for understanding simple stories on the basis of 'scripts' which outline sequences of events or actions to be expected in particular situations — e. g. the 'restaurant script' outlines the normal action-sequence of calling the waiter, ordering food, being served, eating the food, getting the bill, and paying it. One output of the story-understander is a summary of the story extracted and condensed from the full semantic representation (i. e. a kind of 'macrostructural' output). However, it is now proposed by Schank et al. (1980) that equivalent summaries can be produced by text parsers which do not attempt to understand everything in a text and do not need a complete semantic representation. An experimental program FRUMP (DeJong, 1979) works from 'sketchy scripts' of typical newspaper stories (kidnaps, acts of terrorism, diplomatic negotiations, etc.); it skims through texts looking only for words signalling a known 'script', from which it is able to predict or expect the occurrence of other words or phrases and so build up the outline of the story, i. e. a 'summary' of the newspaper report. FRUMP is therefore only 'interested' in and only interprets those parts of the text which relate directly to elements of a 'sketchy script', the rest of the text is ignored.

Text scanning or 'skimming' is pervasive in many areas of information retrieval: users do not, in general, read in full everything put before them, they scan texts to decide which documents are relevant or worth reading; likewise, indexers and abstractors rarely read the full texts to documents, they

scan them for words and phrases indicating general content. Yet 'scanning' has not attracted widespread attention. However, one aspect has been investigated by Keen (1977) in his study, under controlled experimental conditions, of the psychological and (in part) linguistic processes involved in the consultation of printed subject indexes. Obviously, index users do not read every entry; but how are they able to locate entries with the terms they seek? how do they spot entries with related meanings but with terms they had not thought of?

Although the research of DeJong and Schank on text scanning is clearly important, its direct relevance to information retrieval will remain marginal until more work is done on analysing non-narrative texts. Information retrieval is concerned primarily with expository and descriptive texts (scientific papers, research reports, scholarly articles, etc.) which exhibit different structural features from those found in stories and other narrative texts; indeed, it appears that they reflect in their text structure the hypothesis-testing and problem-solving frameworks of the research process itself (Gopnik, 1972; Hutchins, 1977). Some insight may come from text-linguistic analyses of scientific discourse.

### *5. Formatting*

Structured databases of the kinds found in question-answering systems (section 2.3 above) are generally compiled in the same way as their non-computerized printed equivalents, i. e. by human effort. The auxiliary processes of sorting, filing and indexing may be automated to varying degrees but not, in general, the basic activities of restructuring text material into formats suitable for searching. As for the intellectual aspects of collating and organising information and data, almost as little is known about them as is known about the psychological processes of indexing and abstracting.

5.1. There has been, however, considerable research activity on the problems of representing knowledge in databases, particularly in the field of artificial intelligence (cf. Findler, 1979). Although some of this research has involved the representation of information extracted from texts, there has been little work on the problems of handling substantial volumes of textual material or of integrating information from a variety of texts. In this respect, research at SRI International could be important: one project involves the incorporation of textual material from different sources into a single database, the 'Polytext' project (Walker, 1981) the other involves the conversion of a large database at present in text format into a semantic representation suitable for interrogation by a LIFER-type question-answering system (Hobbs, et al., 1982). The database selected is the Hepatitis Knowledge Base (Bernstein, et al., 1980), a computer-based textbook representing the current consensus of

experts in this medical field, compiled and continuously updated by computer conferencing techniques.

5.2. The most substantial research so far on converting natural language material into 'information formats' is the work of Sager and her colleagues (Sager, 1975, 1978, 1981). The team has developed a generalised parser — the linguistic string parser, deriving from Zellig Harris' well-known work in mathematical linguistics — which has been applied to the analysis and representation of medical records in a large database. The automatic 'formatting' of the text takes place in two stages: first, the parser produces a linguistic string analysis of sentences (a phrase-structure, dependency representation which does not distort the original linear sequence); and then, each parsing is segmented into semantic categories appropriate to the content of the text type in question. For example, the following medical record would be 'formatted' as in figure 4:

Patient first had sickle cell anemia diagnosed at age 2 when he complained of leg pain. He was worked up and diagnosis was made. He was asymptomatic until age 5 when he was admitted to Bellevue Hospital with chest pains. He was hospitalized for a month and released.

Information formatted in this way can be used not only as databases for fact-retrieval (Grishman and Hirschman, 1978) but also as data sources for further statistical analyses and as foundations for 'expert systems' (cf. Bramer, 1981).

Successful formatting depends on the validity of the initial analysis of the 'sublanguage' of the texts to be handled. In any discipline there are semantic constraints on the acceptability of certain statements. In cell biology, for example, the statement *the ion crosses the membrane* would be an acceptable proposition (whether true or false in a particular instance), whereas *the membrane crosses the ion* would be rejected as nonsense. Such observations lead to the establishment of 'sublanguage grammars' consisting of subject-specific classifications of vocabulary (e.g. noun-classes for digitalis texts such as 'cations', 'enzymes', 'cells', 'proteins') and subject-specific syntactic rules (e.g. elementary propositions of the form  $N_{ion} V_{move} N_{cell}$ , and causal structures of the form  $N_{drug} V_{affect} (N V N)$ , etc.) These sublanguage categories and structures determine rules of the second stage of analysis and the headings of information formats (i.e. V-MD, V-PART, NORM, SIGN/SYMPT in figure 4.)

5.3. The concept of a 'sublanguage grammar' is obviously closely related to that of a 'semantic grammar' in question-answering systems (section 2.3 above). Many interests converge in the investigation of 'sublanguages', making it a distinct field of research within the general disciplines of text linguistics and computational linguistics (Kittredge and Lehrberger, 1982); information science could also contribute with statistical methods for the automatic classification and differentiation of texts according to subject

CONJ	PATIENT	TREATMENT		PATIENT STATE					TIME			
		INST	V-MD	V-PAT	BODY PART	NORM	SIGN/ SYMPT	DIAGNOSIS	P 1	P 2	REF. PT	
	patient		first had diagnosed							at		age 2 yrs
when	he			complained of	leg		pain					(age 2 yrs)
	he		was worked up									
and			diagnosis was made									
	he			was		asymptomatic				until		age 5 yrs
when	he	Bellevue Hospital	was admitted to	with	chest		pains					(age 5 yrs)
	he		was hospitalized								for a month	
and	(he)		(was) released									

Fig. 4: Information format

content (e.g. in automatic indexing; section 3.2 above). However, it is more probable that 'sublanguage' research will eventually revive interest in linguistics-based approaches to automatic indexing and abstracting.

### *6. Evaluating*

The ultimate test of an information retrieval system, whether a bibliographic system or a fact-retrieval system, is its effectiveness in satisfying the information needs of its users. The future role of text analysis procedures in information retrieval can be put in simple terms: can text analysis improve system performance? and, if so, will the improvement be sufficient to justify any additional costs? However, the questions are not at all easy to answer. There are many complex factors involved when users of systems assess the value of the information provided. First, the documents or data retrieved must be interpreted (read and understood), the particular information required must be extracted, and then it must be 'applied' in solving the problem (or anomaly) which prompted the original request. In the case of bibliographic systems there are additional complexities in that users have to decide which documents are likely to be worth reading on the basis of their titles, abstracts and index entries — and these can easily be misleading or misunderstood. There are numerous occasions for mistaken assessments: unfamiliarity with the subject matter, unfamiliarity with the information system and its indexing/abstracting policy, failure to recognise the 'relevance' of texts to actual needs, etc. The problem of 'relevance' has indeed engaged researchers for many years (Saracevic, 1975); and rightly so, as relevance assessment is at the heart of system evaluation in information retrieval. What is being assessed: relevance to a request, to a search formulation, to a supposed 'information need', or to an ultimate 'solution' of a problem? relevance of a document citation, or of a document text, or of a document's contents? relevance as seen by the user, or by the information supplier, or by an independent judge?

With such a complexity of factors it is not surprising that most researchers are convinced that retrieval from bibliographic systems will long remain essentially a probabilistic process. In this context, detailed analysis and representation of document texts is likely to be always of less significance than interactive searching and feedback techniques for helping users to find the documents they think might satisfy their needs. There is also the question of scale: bibliographic systems are necessarily concerned with the gross characterisation of documents within large collections; there appears to be little justification for the detailed semantic analysis of fact-retrieval and question-answering systems. The latter are generally designed for relatively narrow subject domains and for access by expert users requiring precise information. It is research on these systems that most advances in text understanding and text representation have been made in the field of

information retrieval, and this is likely to continue in the future. However, it is hoped that this overview has shown that text processing of many kinds is an integral part of all information retrieval systems and that research on the understanding of these processes (whether it leads to improved systems, or to computerization, or not) may well contribute something of value in the general field of discourse analysis and text linguistics.

#### *Acknowledgement*

The author wishes to thank in particular Dr. Karen Sparck Jones for her invaluable suggestions and comments on an earlier version of this paper.

#### *Bibliography*

(Notes: \* = survey or review article)

ARIST = Annual Review of Information Science and Technology

JASIS = Journal of the American Society for Information Science)

- Barnes, C. I., Constantini, L. and Perschke, S. (1978). Automatic indexing using the SLC-II system. *Information Processing and Management* 19, 107–119
- \*Bates, M. J. (1981). Search techniques. *ARIST* 16, 139–169
- Baxendale, P. B. (1958). Machine-made index for technical literature – an experiment. *IBM Journal of Research and Development* 2, 354–361
- Beaugrande, R. de and Dressler, W. (1981). *Introduction to text linguistics*. London: Longman.
- \*Becker, D. (1981). Automated language processing. *ARIST* 16, 113–138
- Belkin, N. J. and Oddy, R. N. (1979). Design study for an Anomalous State of Knowledge based information retrieval system. Birmingham, University of Aston: Computer Centre. (British Library Research and Development Report no. 5547)
- Bely, N., Borillo, A., Virbel, J. and Siot-Decauville, N. (1970). *Procédures d'analyse sémantique appliquées à la documentation scientifique*. Paris: Gauthier.
- \*Bernier, C. I. (1968). Abstracts and abstracting. In *Encyclopedia of Library and Information Science*, vol. 1, 16–38. New York: Dekker.
- Bernier, C. I. (1970). Terse literatures, 1: Terse conclusions. *JASIS* 21, 316–319
- Bernstein, L. M., Siegel, E. R. and Goldstein, C. M. (1980). The Hepatitis Knowledge Base: a prototype information system. *Annals of Internal Medicine* 93, 169–222
- Bobrow, D. G., Kaplan, R. M., Kay, M., Norman, D. A., Thompson, H. and Winograd, T. (1977). GUS: a frame-driven dialog system. *Artificial Intelligence* 8, 155–173
- \*Bramer, M. A. (1981). A survey and critical review of expert systems research. In *Information Technology for the Eighties*, R. D. Parslow (ed.), 486–515. London: Heyden.
- \*Damerau, F. J. (1976). Automated language processing. *ARIST* 11, 107–161
- Daneš, F. (1974). Functional sentence perspective and the organization of text. In *Papers on Functional Sentence Perspective*, F. Daneš (ed.), 106–128. The Hague: Mouton.
- DeJong, G. (1979). Prediction and substantiation: two processes that comprise understanding. In *IJCAI-79: Proceedings of the Sixth International Joint Conference on Artificial Intelligence, Tokyo 1979*, 217–222. Stanford, Ca.: Stanford Univ.
- Dijk, T. A. van (1977). Complex semantic information processing. In *Natural Language in Information Science*, D. E. Walker (ed.), 127–163. Stockholm: Skriptor.
- Dijk, T. A. van (1980). *Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition*. Hillsdale, N.J.: Erlbaum.

- Doszkoacs, T. E. and Rapp, B. A. (1979). Searching MEDLINE in English: a prototype user interface with natural language query, ranked output, and relevance feedback. In *Information Choices and Policies: Proceedings of the ASIS Annual Meeting 1979*, 131–137. White Plains, N.Y.: Knowledge Industry Publ.
- Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the ACM* 16, 264–285
- Epstein, M. N. and Walker, D. E. (1978). Natural language access to a melanoma database. In *Proceedings of the Second Annual Symposium on Computer Applications in Medical Care*, 320–325. New York: IEEE.
- \*Findler, N. V. (1979), ed. *Associative networks: representation and use of knowledge by computers*. New York: Academic Press.
- Gardin, J. C. (1973). Document analysis and linguistic theory. *Journal of Documentation* 29, 137–168
- Gopnik, M. (1972). *Linguistic structures in scientific texts*. The Hague: Mouton.
- Grishman, H. and Hirschman, L. (1978). Question answering from natural language medical data bases. *Artificial Intelligence* 11, 25–43
- Harris, L. R. (1977). User oriented data base query with the ROBOT natural language query system. *International Journal of Man-Machine Studies* 9, 697–713
- Harter, S. P. (1978). Statistical approaches to automatic indexing. *Drexel Library Quarterly* 14, 57–74
- Hendrix, G. G., Sacerdoti, E. D., Sagalowicz, D. and Slocum, J. (1978). Developing a natural language interface to complex data. *ACM Transactions on Database Systems* 3, 105–147
- Hillman, D. J. (1968). Negotiation of inquiries in an on-line retrieval system. *Information Storage and Retrieval* 4, 219–238
- Hillman, D. J. (1973). Customized user services via interactions with LEADERMART. *Information Storage and Retrieval* 9, 587–596
- Hobbs, J. R., Walker, D. E. and Amsler, R. A. (1982). Natural language access to structured text. In *COLING 82*, J. Horecky (ed.), 127–132. Amsterdam: North-Holland Publ. Co.
- Hutchins, W. J. (1975). *Languages of indexing and classification: a linguistic study of structures and functions*. Stevenage: Peregrinus.
- Hutchins, W. J. (1977). On the structure of scientific texts. *UEA Papers in Linguistics* 5, 18–39
- Hutchins, W. J. (1978). The concept of 'aboutness' in subject indexing. *Aslib Proceedings* 30, 172–181
- Keen, E. M. (1977). On the processing of printed subject index entries during searching. *Journal of Documentation* 33, 266–276
- Kittredge, R. and Lehrberger, J. (1982), eds. *Sublanguage: studies of language in restricted semantic domains*. Berlin: de Gruyter.
- Klingbiel, P. H. (1973). Machine-aided indexing of technical literature. *Information Storage and Retrieval* 9, 79–84
- \*Lancaster, F. W. (1968). *Information retrieval systems: characteristics, testing and evaluation*. New York: Wiley
- Lancaster, F. W. (1977). Information science. In *Natural Language in Information Science*, D. E. Walker (ed.), 19–43. Stockholm: Skriptor.
- Larson, S. (1977). On-line systems for legal research. *Online* 1(3), 10–14
- \*Larson, S. E. and Williams, M. E. (1980). Computer assisted legal research. *ARIST* 15, 251–286
- Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2, 159–165
- Mathis, B. A., Rush, J. E. and Young, C. E. (1973). Improvement of automatic abstracts by the use of structural analysis. *JASIS* 24, 101–109
- \*Montgomery, C. A. (1972). Linguistics and informations science. *JASIS* 23, 195–219
- O'Connor, J. (1975). Retrieval of answer-sentences and answer-figures from papers by text searching. *Information Processing and Management* 11, 155–164
- O'Connor, J. (1980). Answer-passage retrieval by text searching. *JASIS* 31, 227–239

- \*Petrick, S. R. (1976). On natural language based computer systems. *IBM Journal of Research and Development* 20, 314–325
- Popper, K. R. (1972). *Objective knowledge: an evolutionary approach*. Oxford: Clarendon Press.
- \*Rijsbergen, C. J. van (1979). *Information retrieval*. 2nd ed. London: Butterworths.
- Rush, J. E., Salvador, R. and Zamora, A. (1971). Automatic abstracting and indexing, II: Production of indicative abstracts by application of contextual inference and syntactic coherence criteria. *JASIS* 22, 260–274
- Sager, N. (1975). Sublanguage grammars in science information processing. *JASIS* 26, 10–16
- Sager, N. (1978). Natural language information formatting: the automatic conversion of texts to a structured data base. *Advances in Computers* 17, 89–162
- Sager, N. (1981). *Natural language information processing: a computer grammar of English and its applications*. Reading, Mass.: Addison-Wesley.
- Salton, G. (1968). *Automatic information organization and retrieval*. New York: McGraw-Hill.
- \*Saracevic, T. (1975). Relevance: a review of and a framework for the thinking on the notion in information science. *JASIS* 26, 321–343
- Schank, R. C. (1975). *Conceptual information processing*. Amsterdam: North-Holland Publ. Co.
- Schank, R. C., Lebowitz, M. and Birnbaum, L. (1980). An integrated understander. *American Journal of Computational Linguistics* 6, 13–30
- \*Smith, L. C. (1980). Artificial intelligence applications in information systems. *ARIST* 15, 67–105
- Soergel, D. (1974). *Indexing languages and thesauri: construction and maintenance*. Los Angeles: Melville.
- Sparck Jones, K. (1971). *Automatic keyword classification for information retrieval*. London: Butterworth.
- \*Sparck Jones, K. (1974). Automatic indexing. *Journal of Documentation* 30, 393–432
- \*Sparck Jones, K. and Kay, M. (1973). *Linguistics and information science*. New York: Academic Press.
- Taylor, R. S. (1968). Question-negotiation and information seeking in libraries. *College and Research Libraries* 29, 178–194
- \*Walker, D. E. (1977), ed. *Natural language in information science*. Stockholm: Skriptor.
- \*Walker, D. E. (1981). The organization and use of information: contributions of information science, computational linguistics and artificial intelligence. *JASIS* 32, 347–363
- Waltz, D. L. (1978). An English language question answering system for a large relational database. *Communications of the ACM* 21, 526–539
- Weil, B. H. (1970). Standards for writing abstracts. *JASIS* 21, 351–357
- Weischedel, R. M. and Black, J. E. (1980). Responding intelligently to unparseable inputs. *American Journal of Computational Linguistics* 6, 97–109
- Woods, W. A. (1977). Lunar rocks in natural English: explorations in natural language question answering. In *Linguistic Structures Processing*, A. Zampolli (ed.), 521–569. Amsterdam: North-Holland Publ. Co.
- Woods, W. A. (1978). Semantics and quantification in natural language question answering. *Advances in Computers* 17, 1–87
- Ziman, J. (1968). *Public knowledge: the social dimension of science*. Cambridge: Cambridge Univ. Press.