**Jean-Guy Meunier***
**Suzanne Bertrand-Gastaldy****
**Hermel Lebel**

# A Call for Enhanced Representation of Content as a Means of Improving Online Full-Text Retrieval

Meunier, J.-G., Bertrand-Gastaldy, S., Lebel, H.: **A call for enhanced representation of content as a means of improving online full-text retrieval.**
Int. Classif. 14 (1987) No. 1, p. 2–10, 81 refs.

Given the phenomena of growth and diversification which affect both text databases and their users, it is essential to reflect on the nature of textual information and its representation within the very particular framework of interactive retrieval systems. The latter aim to correlate two types of conceptual structures, that of the user and that of the text, by way of a third structure – the interface. A typology of levels of representation is proposed (typographical, lexical, statistical, linguistic, semiotic, and pragmatic). These representations, obtained by means of a multiplicity of strategies (intra-sentence, intratextual, intertextual) applied to different units of information and interrelated, render the interaction between diverse users and the database more flexible and more adaptable.

*(Original abstract, translated from the French by Peggy Warren.)*

## Introduction

We are interested in the computerized analysis and the representation of text databases in interactive information systems as a means of facilitating information retrieval.

Our research is situated within a rapidly evolving context. In effect, as the demand for and the production of information continues to grow, we see an increase in and a diversification of databases. Several factors contribute to the increase of magnetically stored information:

– a diversification of computerized tools for the creation of these databases: minicomputers, microcomputers, videotext, word-processors (Ingwersen, 1984; Le Loarer, 1986; Normier, 1985);
– a reduction of costs and an increase in storage capacities;
– an integration of the phases of production and of diffusion of databases by large distributors who have begun to produce them as well as by specialized organizations wishing to market their products (Neufeld and Cornog, 1983).

Likewise, the reduction in the price of microcomputers and end user retrieval systems (Janke, 1984; Nicholas and Harman, 1985) has led to a diversification of users with widely varying expectations (Ostrum and Yoder, 1985; Trenner and Buxton, 1985).

Natural language databases (abstract and full-text) have not been immune to these tendencies and their number continues to grow, the latter supplanting the former (Boumans, 1985). This development will become more pronounced with the ever growing number of newspaper articles, magazines, books, encyclopedias, press releases, legal texts, etc., available on-line.

We have only begun to witness the multiplicity of possible uses of databases, as well as the complexity of the task facing the user obliged to consult, for a single project, several databases in different fields containing different types of data, via different host systems including the user's own company (Peterson Holland, 1985).

In the face of such fervent activity, one is struck by how slowly on-line information systems are evolving.

The modes of analysis, retrieval (Peterson Holland, 1985), and display (Tenopir, 1985a) of text information are not significantly different from their counterparts in bibliographical databases. The content of the texts continues to be represented by means of thesauri and classification schemes, thereby giving, according to some, a fixed and monolithic external perspective (Streatfield, 1983), the sole internal perspective often consisting of the inverted file of character strings contained in the base proper. The modes of interrogation are approximately the same, and those specifically conceived for full language are often too complicated for the casual user (Conger, 1984). The potential of full-text databases thus threatens to remain relatively unexploited (Ingwersen, 1984), possibly resulting in a disaffection of the public in the short run.

Judging from preliminary evaluations (Blair, 1986; Tenopir, 1985a), the outcome of research on full-text databases has not been very satisfactory, in that the level of recall and precision is rather low. Still, they remain indispensable for finding documents that no other element – title, controlled vocabulary or abstract – can locate.

The importance of change, the extent of the diversification phenomenon (affecting every aspect of the text databases themselves as much as the users), the costs incurred by continuing problems in retrieval, all prompt a search for solutions other than further elaborating traditional methods of analysis and retrieval or multiplying the power and capacity of computers.

We feel it is important to reflect on the nature of textual information and the manner in which it is represented in light of research in areas as diverse as linguistics, semiotics, psycho-linguistics, artificial intelligence, philosophy, cognitive psychology, and learning theory. First, however, we believe that one must reformulate the theoretical framework of which every information system is a part, and consider the original aspects of the communicatory structure – namely, interactive systems of text databases.

* Université du Québec à Montréal, Département de Philosophie.
** Université de Montréal, Ecole de bibliothéconomie et des sciences de l'information.
Modified version of a paper given in French for the "Third FID/CR Regional Meeting", Montréal, Sept. 13, 1986. Transl. by Peggy Warren and Arthur Henk.

## 1. The Communicatory Aspect of Interactive Full-Text Information Systems

The communication between user and machine in an interactive natural-language information system consists of a *complex interrelation among several elements* aiming at correlating two frames of reference – the cognitive structure of the user and the information structure of the text database – by means of a third intermediary structure, that of computerized interface procedures (in the case of an individual search) and occasionally a fourth structure, that of the librarian (in the case of an assisted search).

> (. . .) interactions of humans with one another, with the physical world and with themselves are always mediated by their states of knowledge about themselves and about that with which or whom they interact. (. . .) we look at the IR[1] situation as a recipient-controlled communication system, aimed at resolving the expressed information needs of humans, primarily via texts produced by other human beings. (Belkin et al., 1982, p. 85).

*Each of the interacting structures displays different characteristics:*

(a) on the one hand, the human being – intelligent, adaptable, flexible, able to reason but incapable of memorizing and processing great numbers of texts;

(b) on the other, the full-text database – a reflection of the cognitive structures of several authors having fixed in the texts their cognitive states at a given moment. This structure may seem, at first view, to be easier to apprehend than the preceding structure. However, the semantic richness of the text database is inexhaustible and can be represented in a number of ways, using various characterizations and strategies of analysis.

(c) between the two, the interactive system – by turns receiver of the user's question and transmitter of a response, capable of memorizing and processing great quantities of data, but having a relatively simple and rigid structure despite efforts, based on research in artificial intelligence, to make it imitate human reasoning by instilling in it a certain quantity of knowledge.

The user does not directly scan the texts, at least not at first, in contrast to what occurs in the activity of reading. The machine scans the texts for the user, within the limits of comprehension and interpretation inherent in the machine. This fact invites us to redefine the interaction of the user not vis-a-vis the machine, but via the machine towards the texts.

The presence of a human being, endowed with a cognitive structure, in the role of information seeker is extremely important. It is, in effect, only through the intervention of a cognitive agent that the data (e.g., words or complete texts) can be transformed into information, and these data must be capable of being connected to some element in the cognitive structure of the individual.

This approach, inspired by cognitive theory (Anderson, 1976; Kintsch, 1977; Neisser, 1967) and proposed by several specialists in information science (Belkin, Ingwersen, Wormell and Pylyshyn, among others) seems best suited to our problem.

*Cognitive structure is characterized by complexity:*

(a) In effect, *all individuals possess their own cognitive structures* resulting from their experiences, their education, their culture, their background (Ingerwersen, 1982), and their knowledge of the different elements of the search.

Indeed, a single information system is used by *numerous individuals* (Janke, 1984 has asserted that the clientele will hereafter be as varied as humankind). The concept of user-friendliness thus becomes a very personal one.

(b) At first, the need for information is often *vague and general.* The questions directed at the system reflect a lack, an insufficiency of knowledge, for Taylor, a "Conscious Need," and for Belkin and Ingwersen, an ASK ["Anomalous State of Knowledge"] (Ingwersen, 1984).

(c) In short, the users, for the most part, do not know exactly what they are looking for. Furthermore, they are often *unaware of what is available*, that is to say, *the actual contents of the database* or how to find information, for the representations of the content of the database are infrequent and relatively rudimentary.

(d) Moreover, the cognitive structure of each individual, taken separately, is not fixed. It *evolves* according to the responses obtained in the course of the communication. But since the user's needs become focussed as the investigation of the database progresses, one may imagine that there comes a moment when the initial representations of the database are no longer satisfactory.

Thus to the purely synchronic dimension of the communication is added another, very important one: the *diachronic* dimension. In effect, the system should not only adapt itself to the user, it should also adapt itself to the user's change of cognitive state as time passes, by bringing into play, from among its different possibilities of processing, the one which best responds to the need expressed at a given moment. The user, after posing a question, receives a response which modifies his or her state of knowledge. This may lead to the posing of another question which, in turn, produces a further change, and so forth.

This manner of conceiving of the communicatory situation of text retrieval leads us to favour an approach offering the user *several diverse "portraits" of the database* (that is, diverse representations of the information contained there). These portraits may correspond to the type of question (functional aspect) and especially to the user's state of knowledge (cognitive aspect). Furthermore, we believe that the strategies of analysis for retrieval performed by the user in constant interaction with the database should be different from those employed in artificial intelligence systems simulating understanding or even translation, or in automatically formatted systems. The user would doubtless be relieved to be helped in seeking information, but certainly would not stand for being completely replaced by the computer; the user would rather keep an active role.

While one notes a growing attention to the cognitive sciences and to psychological ergonomics in the context of information systems, one is obliged to note that the majority of current systems are neither flexible nor adaptable.

We shall thus examine, in the following discussion, whether it is possible to find, in text representation, at

least a partial response to this need for flexibility and adaptability.

## 2. The Representation of Text Databases

### 2.1 The complexity of the representation of text information

In the memory of the computer, one finds the text database, itself possessed of a structure which, while more coherent than that of the individual in search of information, is no less complex.

*This complexity is due to several factors:*

(a) Computerized content analysis brings into play *several concepts of texts.*

The user searches among the subjects the particular topics transcribed into the memory of the computer in the form of electronic texts (Text E). These electronic texts are copies of material texts (Text M) – periodical articles, for example – which themselves have a discursive content created by the author (Text D) (Meunier and Lepage, 1982).

Furthermore, a text is not only a succession of insolated, quantifiable lexical units nor an unordered succession of sentences; it constitutes a complex organic entity whose meaning evolves from the integration of successive graphic, linguistic, and semantic units of increasing size.

The strategies of text analysis may thus generate various representations of the organization of text information.

This hypothesis of *levels of representation of text information* belongs to a theoretical movement recurring in linguistics (Chomsky; Fodor, 1974), in cognitive psychology (Anderson, 1976) in text theory (Beaugrande, 1980, 1984; Van Dijk, 1977), and in artificial intelligence (Charniak, 1972; Schank, 1975; Wilks, 1972).

(b) The database contains hundreds, even thousands of texts written by different authors. On the one hand, the informational value of each text depends, to a certain extent, on the manner in which its content can be differentiated from other texts present in the database. On the other hand, the meaning of each text is in part a function of its presence in the collection of other texts. In another corpus, it would take on a completely different meaning. A good representation should thus take into account this *intertextual dimension.*

(c) *The interrelation of different representations obtained with respect to the original text* adds to the complexity. Each representation may or may not derive from a previous representation. Each text submitted to a strategy of computerized analysis is an original text which is transformed into a new text, which may itself serve as an original text for a subsequent strategy of analysis. To each of these texts one may associate a representation – which is, in this perspective, an organization of the information contained in the text. Since the representations obtained from a given original text vary as a function of the units of processing and strategies of analysis selected, one can easily imagine that the search strategies and the responses obtained also vary.

Thus through a succession of analyses, one goes from a text database in natural language to a formatted database. For example, from a group of texts one may extract either a list of words organized in an inverted file or an automatically formatted database, ready to respond to the type of specific questions that one poses in question-answer systems, because the text information will have been integrally conserved, then categorized, after a reduction of the variety and complexity of its modes of expression (Sager, 1981). Between these two extremes, there obviously exists a whole range of possible representations resulting from either an increase or a reduction of the original text data[2].

Let us examine some of these possible representations in terms of a provisional typology that we have established and illustrated with several examples regarding the type of representation, the type of strategy used to produce it, and the type of information one may draw from it[3].

### 2.2 Typology of text representations

*(1) Representations constructed from certain typographical properties*

The first representation applied to Text M for recording in the memory of a computer is that comprising the ensemble of its typographical characters and their transcription into electronic form. This representation rests on strategies of character reading and recognition which seem uncomplicated but which actually involve a number of problems such as lack of uniformity and multiplicity of typefaces (boldface, italics, underlining, etc.). The interface system responds to questions regarding the correlation between a given character and one of those it has memorized. If, at this initial stage of analysis, one limits the characteristics retained, one runs the risk of eliminating subsequent analysis strategies. Thus, the absence of French diacritical marks prevents one from arriving at the lexical level, and the failure to identify acronyms and languages limits linguistic analysis, etc.

This representation is thus an indispensible step for the extraction of vocabulary by subsequent algorithms.

*(2) Lexical representations*

Lexical representation in its most reduced form consists of an alphabetical list of character strings. Strategies of analysis carried out without human intervention have several advantages: rapid, exhaustive, economical and "objective", they render immense text databases immediately accessible in their entirety. Supplemented by location numbers, this representation leads to the production of concordances, permuted indexes or inverted files. It is also the avenue to several statistical, syntactic, and semantic strategies of analysis.

Lexical representations nonetheless present a certain number of disadvantages resulting from the very rudimentary character of the information that they can provide. In effect, even if the original text is in natural language, the transformed text has few points in common with this natural language. All that remains is graphical signs without meaning and without paradigmatic or syntactical dimensions. Relations based on the signs themselves (homonymy, polysemy) are absent, as

are relations based on their meanings (synonymy, equivalency classes, hierarchy, relations of semantic association) and relations based on the reference. Elements of meaning constituting the content of the texts are neglected. One does not take into account the organization of the text, the context, or the network of relationships among its terms. Thus, we can assert that this type of processing, far from reducing the variety of expressions in the original text, accentuates this variety, running counter to the objectives of analysis in an information system and constituting what might almost be called "anti-documentation" (Long, 1980, p. 105).

In fact, this type of representation is intended to assist the user in the rapid isolation of subsets that are very small in relation to the database as a whole, leaving to the user the task of scanning the corpus of texts in their entirety of passages of text retrieved. The user's mind thus plays its usual role in interpreting the graphical signs it reads in the light of its knowledge of the language, the area in question, and the particular problem to be resolved.

Hence the importance of the cognitive agent and its role as "stand-in" when faced with a very summary representation. Hence also the claim by some that this type of representation is sufficient, since it gives reasonably good results in retrieval. But the user disposes of a series of strategies permitting the simulation of higher levels of analysis (syntactical and also semantic) on units of information not characterized linguistically, and the reintroduction – in a very impoverished form, to be sure – of certain of the connections among words found in the original text:

- the display of the context allows the user to evaluate the meaning of the word and the relevance of the passage with respect to the answer sought and to modify the search strategy accordingly, notably by means of exclusion operators which improve the focus of the search by eliminating non-relevant contexts.[4]
- moreover, the use of adjacency operators in the search strategy offers the possibility of reconstituting a simulacrum of phrases. This is still a far cry from the precision of natural syntax, since such operators do not describe these connections among words in the original text, nor do they permit the introduction of prepositions into the search strategy. Furthermore, false relations among the words persist.
- finally, truncation allows one to exercise a semblance of semantic reordering around a single radical, still with a good deal of imprecision since it remains a reordering based on character recognition.

The manipulation of certain operators is nonetheless very difficult for non-experts; decidedly too much is demanded of these casual users, with the result that they generally limit themselves to the principal operators, to the detriment of the optimal exploitation of natural language texts (Ingwersen, 1984).

### *(3) Representations obtained from certain statistical characteristics*

Some representations derive from numerical characteristics of occurrences, others, from operations of automatic classification and multivaried analysis of co-occurrences. Generally applied to basic lexical representations of character strings, the former can reveal, in a highly condensed form (Meunier et al., 1976) certain sets of themes but tell nothing about the interdependence of lexical forms taken separately and do not permit the identification of structures or networks of significations that might reveal implied semantic dimensions.

Despite its elementary nature, the first type of representation – rapid and inexpensive – is frequently employed in documentation as an aid to retrieval in the case of very large corpuses. It has begun to be made more adaptable and to be used for more sophisticated explorations; in the future, more systems will offer the possibility of displaying the occurrences in a subset of the database (several texts designated in advance by the user with the aid of Boolean operators) or in a sub-set of a single text (title, paragraph). Thus, even using simple representations, the cognitive agent can enjoy a certain amount of flexibility:

> Zoom is designed to analyse the frequency of single words, phrases or codes appearing in a selected set of references. Up to a maximum of 200 records can be analyzed. (Ingwersen, 1984, p. 481).
> Zoom was invented and applied to the ESA-Quest search language in order to improve and support the casual user's search performance. (Ibid., p. 481).

Finally, software such as SIRE, or its adaptation EDIBASE, display in graphic form the number of documents retrieved during a search, by decreasing order of relevance, according to the number of occurrences in each document of the word requested.

The second form of representation, for its part, permits the detection of certain networks of information that even a fastidious study of concordances would not have revealed (Lebel, 1985). This type of representation rests on the psycho-linguistic principle (Lyons, 1978; Osgood, 1959) whereby the ensemble of a word's contexts contributes, in some measure, to determining the word's contexts contributes, in some measure, to determining the word's meaning – in any case, it gives information as to the perspective from which the concept is considered (Ford, 1983).

Due to the simplicity and rapidity of processing, computerized classification is frequently used in library and information science to index texts by automatic means and to reorganize the database by bringing together texts of similar content. Research in this area began with Doyle, (1962) Needham and Sparck Jones, (1964). The work of Salton, too, is well known, and this type of research continues to proliferate. In the realm of content analysis, one might also cite Fisher and Langley, 1985; Lebovitz, 1983; and Michalski, 1980. However, there exists as yet relatively little work on obtaining representations from the entire database by means of what has sometimes been called a search thesaurus; AID of Doszkocs (1979) is a notable exception. The advantage of this tool lies in its adaptability. Rearrangement of terms takes place on demand, in the desired subset of the database. But statistical analyses of co-occurrences of words, within documents and in the database as a whole, should spread rapidly. The project ASK aims represent-

ing not only the texts of the database, but also the questions formulated by the users:

> There is no question that this sort of representation of a state of knowledge (or of the information structure underlying a text) is simplistic and naive, if one is attempting to obtain detailed representations for such purposes as natural language understanding, machine translation or retrieval from memory. On the other hand, it has the advantages of being fairly easily determined and reasonably machine-manipulable, important considerations in an IR context, where one needs to represent actual information needs and to manipulate large amounts of data. (Belkin et al., 1982, p. 68).

The disadvantage is that these representations in no way specify the nature of the connections between words. This interpretation is left to the searcher. These methods of representation moreover imply a selection of certain lexical forms and pose considerable theoretical problems.

### *(4) Representations constructed on the basis of linguistic characteristics*

In library and information science, statistical representations most often take a typographical representation as their point of departure. One may, however, add linguistic characterizations to each occurrence of character strings, thereby expanding the previous rudimentary representation by means of strategies of lemmatization, morphological analysis, and syntactic characterization (noun, adjective, verb, etc.).

From the new representation, one can then proceed to a syntactic characterization which associates, with the preceding representation, the positions of each morpheme within a phrase and of each phrase within the sentence. In the current state of research, this syntactic description can be carried out more or less automatically depending on the language. For French, one thinks of the Deredec system's GDS grammar (Plante, 1983).

In information science, the high price of syntactic analysis has always been a source of concern and less expensive expedients have been sought, notably in the commands of the search language, as we have seen above. Yet even if one assumes that a user of an information system is unlikely to be interested in a syntactic representation, it is easy to see that such a representation constitutes a prime example of a superior model (of a semiotic or even statistico-linguistic order, as we shall see below).

The most advanced instance of the use of syntax (Harrissian) is surely the *Linguistic String Project* which, by way of several intermediate strategies and representations (parsing, grammatical regularization, information formatting, normalization), produces an automatic formatting of all the information contained within the text (Sager, 1981).

### *(5) Semiotic representations*

These representations associate, with any given previous representation, information perceptible at the level of the signs themselves.

### *(a) Semantic representations*

*Semantic* representations add this type of information to morphemes or phrases. One thinks of classic systems of content analysis (cf. Stone, 1966, who added to each word of a text a category generated by a "thesaurus" termed political and even psychological). Systems of artificial intelligence associate, with each word and especially with each sentence, a representation of its conceptual structure (Beaugrande, 1980; Schank, 1977; Wilks, 1972) and even of its logico-functional form (Fredericksen, 1977; Jackendoff, 1983).

In most work in library and information science, the entire text (a unit obviously much larger than a word or sentence, but adapted to the size of the database) has been represented by descriptors taken from a thesaurus or by classification numbers. As we know, the conversion of words in natural language into descriptors can be performed by computer. In certain full-text databases, a partial human analysis is carried out to add paradigmatic relationships. This, for example, is the case of NEXIS, which offers a certain control of synonymic forms including: British and American spellings; acronyms of governmental organizations; abbreviations and complete forms; typographical differences between compound words with or without hyphen.

The regrouping of terms by means of semantic relations contained in thesauri differs from the statistical regroupings mentioned earlier. These latter cases, constructed on the basis of co-occurrences, are not founded on relationships of meaning.

> (. . .) whereas the semantical relationships are based solely on the meanings of the terms and hence independent of the "facts" described by those words, the statistical relationships between terms are based solely on the relative frequency with which they appear and hence are based on the nature of the facts described by the documents ((Maron and Kuhns, 1960, p. 225).

Thus the distinction becomes clearer between traditional thesauri and search thesauri based on the co-occurrences of words in the texts. Their complementarity is also better understood. Traditional thesauri organize the concepts of a discipline according to their common characteristics; search thesauri emphasize the angles from which the concepts are being considered in one context or another (Bertrand-Gastaldy, 1984).

### *(b) Textual and intertextual representations*

The overall meaning of a document depends on the interrelation of its various elements at a given level and their integration into the units of the next higher level. Any strategy intended to represent the content of a text must attempt to bring out this unifying framework.

Models attempting to describe this framework are proliferating. Some derive from the theories of artificial intelligence (Schank, 1975; Wilks, 1972), others, from theories of logic (Petöfi, 1979), from cognitive theories (Kintsch, 1974), or from theories of learning (Crothers, 1972; Frederiksen, 1977; Meyer, 1975). Still others stem from semiological theories (Greimas, 1976; Levi-Strauss, 1958; Propp, 1970). They have in common the conception of this textual unity as a *grammar*, that is, a set of rules controlling the unification of content. Hence there appears a new level in the organization of meaning – that of the text rather than that of the individual word.

In information science, there have been a number of attempts to push beyond the limits of the sentence. If

traditional linguistic theories have furnished no satisfactory answer (notably for computerized condensation and indexation), it is, according to a number of researchers, because linguistic theories have been restricted to words within sentences or to sentences considered individually. It is possible that new theories of text grammar will be more relevant for information science in its concern with the message transmitted by the text.

> Informatics must be concerned with text, rather than merely sentence, and must be concerned with semantics within the text. The refusal of theoretical linguistics to deal with this problem has been a great hindrance to fruitful cooperation between the two disciplines (. . .). (Dea and Belkin, 1978, p. 75).

One thinks, for example, of work on the theory of "clause relations" (Dea and Belkin, 1978) and also of the *Functional Sentence Perspective* theory (mentioned above), modified and extended to text, as proposed by Janos (1978).

All the work in this direction attempts to bring out the thematic progressions contained in the text. One thus speaks of "thematic metatext" (Janos, 1978), of "supersyntactic units" (Bondorenko, 1975) useful for automatic condensation and extraction (Maeda, 1981). Thus there appear new segmentations for document analysis based on the relations among sentences (causality: "consequently"; contrast: "however"; illustration: "for example"; specification: "in particular").

By representing the structure of the text hierarchically, some foresee the possibility of an analysis several levels deep (Bondorenko, 1975; Marcyszewski, 1976; Meyer, 1975).

It is by examining each text in a homogeneous corpus that one succeeds in bringing to light certain regularities. Beacco and Darot (1984) have studied how, in the bibliographic abstracts of a database, regularities pertaining to discursive and presentational operations become evident. Heslot (1983), for her part, has examined the marks of presentation in scientific discourse. Although certain systems of document analysis make no explicit reference to the thesis of intertextuality (Foucault, 1969), it is implicit in several studies. It served to establish the categories of information for automatic formatting in the *Linguistic String Project*, and it refers to the theory of sub-languages (Kittredge and Lehrberger, 1982). It is of such importance that some have made it the centre of their research (Courtine, 1981; Pêcheux, 1969).

### *(c) Pragmatic representations*

A text, in addition to being an organically structured significant unit, is an act of communication presenting certain statements. It is a complex discursive unit. Certain representations seek to increase the information of a text with respect to the situational context of the statement: time, place, speakers, illocutionary force, frame of reference, conversational involvement, etc. (Searle, 1985; Wilks, 1985).

As far as we know, this representation does not seem to have received attention in the field of library and information science, probably on account of its complexity.

This survey demonstrates the multiplicity of models for the representation of content and suggests the number and complexity of problems remaining to be solved, among them, the compatibility of different possible representations and their computerization. Some do exist and are used in experimental contexts, while others have not progressed beyond the theoretical level.

## 3. Our Research Program

For our part, we maintain that:

- Several areas of research have contributed to broadening the scale of models for content representation.
- Several specialists in information science have become interested in the text as a unit of information.
- It must be remembered that text representations are complex, multiple, and created from a number of different units of information.
- The cognitive model suggests conceiving of these representations in terms of an iterative process whereby the type of representation is made to correspond to the cognitive state of the user, the two evolving together.

From this point of view, an adequate consideration and conceptualization of full-text interactive retrieval systems should lead us to:

- Validate, specify, and correct, as necessary, the typology of the representations;
- Use this typology as a starting point for establishing, on a theoretical basis, typologies that we wish to perfect;
- Match questions with corresponding typologies in order to determine the order of priority of the representations we wish to offer;
- Avoid excluding any representation at the outset on the basis that users do not ask for it. It is known that users have a tendency, through a process of self-censorship, to ask of a system only what they think it can give them;
- Keep in mind, nonetheless, that certain representations may prove to be useless or unprofitable on account of the presence of a cognitive agent;
- Maintain a balance of quality and cost, on the one hand by limiting complex and expensive strategies to the processing of small sub-groups of the database, after the user has made his question sufficiently specific, and, on the other hand by accepting summary analyses for larger corpuses explored, in the initial stage, by users unfamiliar with the general structure of the database or still unable to clearly define their informational needs;
- Conceive of the representations as superimposed systems, one set within the other. The user would thus have at his disposal a range of representations, from the most general to the most specific, upon which to draw in order to delve into the details. It would operate much like a zoom lens (Bertrand-Gastaldy and Davidson, 1986).
- Seek to organize different analysis programmes into modules. As theoretical research progresses, the

sophistication of the representations should grow from a set of simple strategies, all relatively easy to perfect and applicable in other modules.

We have seen that certain work in information science aims at representing the structure of a text database. We too wish to work in this direction, taking, however, as our point of departure elements of texts to which we will attribute supplementary characteristics: typographical properties (diacritical marks, differentiation of languages, positioning within the text, etc.); linguistic properties (lemma, radical, syntactic category, syntactic role, synonomy in the case of initials); and even semiotic properties (identification of theme and rheme; the search for certain metastructures). The combination of several different types of strategies, all well-mastered, may give rise to representations capable of responding to needs not yet satisfied by full-text retrieval systems. It would seem, in particular, that the use of statistical strategies in conjunction with another type (or types) of strategies might furnish whatever assistance the majority of users might need in the course of their searches. This seems to us a fruitful line of research. It is, moreover, the line suggested several years ago in a different context:

(. . .) The intervention of lexicometry in methodologies using surface processing and incorporating both syntax and hypotheses of a semantic order into the corpus, constitutes an inevitable prolongation of the search process. Exhaustive surface processing gives interesting results and presents certain methodological advantages, but of the many possible lexicometric approaches, it is the most elementary. The correlation, by statistical means, of vocabulary, syntax and actual statements is entirely conceivable; it is obviously tied to progress in linguistic theory. In any case, the era of "frequency reduction" is a revolutionary era for lexicometry. One must also hope for a more systematic connection between quantitative and non-quantitative approaches – the two being too separate, to their mutual disadvantage. (Maingueneau, 1976, p. 45).[6]

We are similarly encouraged along this line by the very authors who have worked on basic lexical units. In fact, Sparck Jones (1975) suggested that the performance measures in automatic retrieval could be applied to a wider range of types of descriptions than the usual ones (keyword or stem lists). Salton (1985) has recently imagined that it might be possible to have recourse to a syntactical analysis of texts from the point of view of systems of questions and answers, a path he had previously dismissed. And Hass Weinberg and Cunningham (1984) suggest a combination of statistical and positional approaches.

As for Hirshman and her team (1975), they have obtained representative regroupings of different facets of a sub-field, pharmacology, by applying an analyzer of co-occurrences to words from texts already categorized by means of syntax and a scientific lexicon.

## Conclusion

The evidence suggests that, in an on-line full-text database, it is as fruitless to exclude the conceptual structure of a text as to ignore the cognitive structure of the user!

Although few information systems take into consideration the structure of the texts, we are convinced that this undertaking becomes more and more urgent in the current context.

With full-text databases proliferating in more varied situations, it is likely that the profitability of more advanced analyses will be assured by the flexibility of their uses. Moreover, the obstacles imposed by the cost and unavailability of software are in the process of disappearing and it is a safe bet that they will fall more rapidly as theoretical research progresses.

Representations of texts should provide tools capable of responding to the needs of different users without restriction.

It would certainly be helpful for the heterogeneous clientele of databases to be presented with several levels of structure inherent in the text database and in each individual text contained in the database, in addition to traditional documentary tools. These structures can be brought out by different strategies (intra-sentence, intratextual, intertextual) applied to different units of information. It is only when one has supplied these different representations that one will be able to observe how different individuals, with different training, at different points in their dialogue with the database, make use of the available representations and how these representations can produce a change in the user's own conceptual representation.

## Notes

1 Information retrieval.

2 The above-mentioned factors belong to the synchronic dimension. The very evolution of the content of the database (the constant addition of new texts and the possible withdrawal of outdated texts) adds a *diachronic* dimension to the problem of representation.

3 The reader will recognise the similarities between this typology and that of levels of automatic processing of natural language established by T.E. Doszkocs ("Natural Language Processing in Information Retrieval," *J. ASIS*, 37, 1986, No. 4, p. 194).

4 Research on the usefulness of context display began early, as witness the works of O'Connor, 1973.

5 In addition to the semiotic form, intertextual representation can take a statistical form (cf. discrimination value, differential term-weighting and measures of similarity in the works of the Salton research group, among others).

6 Translated from the French for the purposes of this paper.

## References:


(1) Anderson, J.R.: Language, memory and thought. Hillsdale, N.J.: L. Erlbaum Associates, 1976. 260 p.

(2) Beacco, J.-C., Darot, M.: Analyses de discours: lectures et expression. Paris: Hachette/Larousse, 1984. 175 p.

(3) Beaugrande, R. de: Text production; toward a science of composition. Norwood N.J.: ABLEX, 1984. 400 p.

(4) Beaugrande, R. de: Text, discourse, and process; toward a multidisciplinary science of texts. Norwood, N.J.: ABLEX, 1980. 351 p.

(5) Belkin, J.J. (and others): Ask for information retrieval: Part II. Results of a design study. J. Doc. 38 (1982) No. 3, p. 145–164.

(6) Bertrand-Gastaldy, S.: Les thésaurus de recherche: des outils pour l'interrogation en vocabulaire libre. Argus 13 (1984) No. 2, p. 51–58.

(7) Bertrand-Gastaldy, S., Davidson, C.H.: Improved design of graphic displays in thesauri – through technology and ergonomics. J. Doc. To be published in December 1986.

(8) Blair, David. C.: Full Text retrieval: evaluation and implications. Int. Classif. 13 (1986) No. 1, p. 18–23.

(9) Bondorenko, G.V.: Study of the text as a hierarchic structure of supersyntactic units. Nauchno-Tekh. Inf., Ser. 2 (1975) No. 8, p. 19–24.
(10) Boumans, J.-M.: Getting the business community online; it's quite a different game. In: 9th International Online Information Meeting, 3–5 December 1985. London, Oxford, Eng.: Learned Information, 1985. 486 p., p. 289–293.
(11) Buntrock, R.E.: Full text and polymer. Datab. 7 (1984) No. 4, p. 112–113.
(12) Charniak, E.: Toward a model of children's story comprehension. Cambridge: MIT diss., 1972.
(13) Conger, L.D.: Types of data bases – some definitions. Datab. 7 (1984) No. 1, p. 94–95.
(14) Courtine, J.J.: Analyse du discours politique. Langages (1981), No. 62, p. 9–128.
(15) Crothers, E.J.: Memory structure and the recall of discourse. In: Freedle, R.O., Carroll, J.B., eds. Language comprehension and the acquisition of Knowledge. New York: Wiley, 1972. 380 p., p. 247–284.
(16) Dea, W., Belkin, N.J.: Beyond the sentence: clause relations and textual analysis. In: Jones, Kevin P., Horsnell, V. Informatics 3; Conference held by Aslib Co-ordinate Indexing Group; 2–4 April 1975; Emmanuel College, Cambridge. London: Aslib, 1978. 137 p., p. 67–84.
(17) Dijk, T. Van: Text and context. London: Longmans, 1977.
(18) Doszkocs, T.E.: AID: An associative interactive dictionary for online bibliographic searching. College Park, MD: University of Maryland, 1979. 124 p.
(19) Doyle, L.B.: Indexing and abstracting by association. Amer. Docum. 13 (1962), p. 378–390.
(20) Fisher, M.D.: Langley O.P. Conceptual clustering. In: Gale, W. Artificial intelligence and statistics: Addisson Wesley, 1985.
(21) Fodor, J.: The psychology of language: an introduction to psycholinguistics and generative grammar. New York: McGraw-Hill, 1974. 537 p.
(22) Ford, N.: Knowledge structures in human and machine information processing; their representation and interaction. Soc. Sc. Inf. Stud. 3 (1983), p. 209–222.
(23) Foucault, M.: L'archéologie du savoir. Paris: Gallimard, 1969. 275 p.
(24) Frederiksen, C.: Semantic processing units. In: Freedle, R. ed. Discourse production and comprehension. Norwood, N.J.: Ablex, 1977. 345 p., p. 57–88.
(25) Greimas, A.J.: Du sens. Paris: Larousse, 1976.
(26) Hass Weinberg., B., Cunningham, J.A.: Word frequency data in full text database searching. In: National online meeting proceedings. April 10–12, 1984, 484 p., p. 425–432.
(27) Heslot, J.: Récit et commentaire dans un article scientifique. DRLAV, Rev. de ling. 29 (1983), p. 133–154.
(28) Hirschman, L. (and others): Grammatically-based automatic word class formation. Inform. Process. & Managem. 11 (1975), p. 39–57.
(29) Ingwersen, P.: Psychological aspects of information retrieval. Soc. Sc. Inf. Stud. 4 (1984), p. 83–95.
(30) Ingwersen, P.: Search procedures in the library analysed from the cognitive point of view. J. Doc. 38 (1982), p. 165–191.
(31) Jackendoff, R.: Semantics and cognition. Cambridge, Mass.: MIT Press, 1983. 283 p.
(32) Janke, R.V.: Online after six: end user searching comes of age. Online 8 (1984) No. 6, p. 15–29.
(33) Janos, J.: Results of an experiment with automatic extracting and the problems of the use of condensed texts in automated information systems. Int. Forum, Inform. & Doc. 3 (1978), p. 13–17.
(34) Kintsch, W.: Memory and cognition. New-York: Wiley, 1977. 490 p.
(35) Kintsch, W.: The representation of meaning in the memory. New York: Wiley, 1974. 279 p.
(36) Kintsch, W., Dijk, T. Van: Toward a model of text comprehension and production. Psychol. Rev. 85 (1978), p. 363–394.
(37) Kittredge, R., Lehrberger, J., eds.: Sublanguage: Studies of language in restricted semantic domains. Berlin, New York: de Gruyter, 1982. 240 p.
(38) Lebel, H.: Problèmes méthodologiques découlant de l'application des méthodes de classification automatique à l'analyse de texte par ordinateur. Montréal: Université du Québec à Montréal, MA Thesis, 1985.
(39) Lebovitz, M.: Generalization from natural language text. Cogn. Sc. 7 (1983) No. 1, P. 1–40.
(40) Le Loarer, P.: Traitement du langage naturel et recherche en ligne. Charenton: ERLI, juin 1986.
(41) Lévi-Strauss, C.: Anthropologie structurale. Paris: Plon, 1958. 452 p.
(42) Long, B.: Linguistique et indexation. Docum. 17 (1980) No. 3, p. 99–106.
(43) Lyons, J.: Eléments de sémantique. Paris: Larousse, 1978. 295 p.
(44) Maeda, T.: An approach toward functional text structure analysis of scientific and technical documents. Inf. Process. & Managem. 17 (1981) No. 6, p. 329–339.
(45) Maingueneau, D.: Initiation aux méthodes de l'analyse du discours; problèmes et perspectives. Paris: Hachette, 1976. 192 p.
(46) Marcyszewski, W.: From the concept of the topic of a sentence to the concept of keyword; remarks on a research program. Nauchno-Tekh. Inf. Ser. 2 (1976) No. 11, p. 18–25.
(47) Maron, M.E., Kuhns, J.L.: On relevance, probabilistic indexing and information retrieval. J. of the ACM 7 (1960), p. 216–244.
(48) Meunier, J.-G. (and others): A system for text and content analysis. Comput. and the Human. 10 (1976) No. 5, p. 281–286.
(49) Meunier, J.-G., Lepage, F.: Formal semantics and computer text processing. Comp & maths with appls 9 (1982) No. 1, p. 83–95.
(50) Meyer, B.J.: The organization of prose and its effects on memory. Amsterdam: North Holland, 1975.
(51) Michalski, R.: Knowledge acquisition through conceptual clustering: a theoretical framework and algorithm for partitioning data into conjunctive concepts. Int. J. of Policy Anal. and Inf. Syst. 4 (1980) No. 3, p. 219–243.
(52) Needham, R.M., Sparck Jones, K.: Keywords and clumps. J. Doc. 20 (1964), p. 5–15.
(53) Neisser, U.: Cognitive psychology. Englewood Cliffs, N.J.: Appleton-Century-Crafts, 1967. 351 p.
(54) Neufeld, L.M., Cornog, M.: Secondary information systems and services. Annu. Rev. of Inf. Sci. and Tech. 18 (1983), p. 151–183.
(55) Nicholas, D., Harman, J.: The end-user: an assessment and review of the literature. Soc. Sci. Inf. stud. 3 (1985), p. 173–184.
(56) Normier, B.: Cas d'applications industrielles de l'analyse du langage naturel. (s.l.): ERLI, October 1985. 11 p.
(57) O'Connor, J.: Text searching retrieval of answer-sentences and other answer passages. J. ASIS 24 (1973) No. 6, p. 445–460.
(58) Osgood, C.E.: The representational model. In: de Sola Pool, I., ed. Trends in content analysis, papers. Urbana: University of Illinois Press, 1959, 244 p., p. 33–89.
(59) Ostrum, G.K., Yoder, D.K.: Chemists as end-user searchers – Training and follow up. In: 9th International Online Information Meeting; 3–5 December 1985. London, Oxford, Eng.: Learned Information, 1985. 486 p., p. 131–140.
(60) Pêcheux, M.: Analyse automatique du discours. Paris: Dunot, 1969. 139 p.
(61) Peterson Holland, M.: ZyIndex: full text retrieval power. Online 9 (1985) No. 4, p. 38–42.
(62) Petöfi, J., ed.: Text vs sentence: Basic questions of text linguistics. Hamburg: Buske, 1979.
(63) Plante, P.: GDSF, une grammaire Déredec des structures de surface du français. Montréal: Service de l'informatique, Université du Québec à Montréal, mai 1983.
(64) Propp, W.: Morphologie du conte. Paris: Seuil, 1970. 254 p.
(65) Pylyshyn, Z.W.: Intelligent database interfaces: a survey of some artificial intelligence applications. London, Ontario, Canada: Centre for Cognitive Science, August 1985. (Cognitive Science Memorandum. COGMEM; 17)
(66) Sager, N.: Natural language information processing: a computer grammar of English and its applications. Reading, Mass.: Addison Wesley, 1981. 399 p.

(69) Schank, R.C.: The role of memory in language processing. In: Cofer, C.N., ed., The structure of human memory. San Francisco, 1975. 213 p., p. 162–189.
(70) Schank, R.C.: Scripts, plans, goals and understanding: an inquiry into human knowledge structures. Hillsdale, N.J.: L. Erlbaum, 1977. 248 p.
(71) Searle, J.R.: Foundations of illocutionary logic. Cambridge, New York: Cambridge University Press, 1985. 227 p.
(72) Sparck Jones, K.: A performance yardstick for test collection. J. Doc. 31 (1975) No. 4, p. 266–272.
(73) Sparck Jones, K. Kay, M.: Linguistics and information Science. London: Academic Press, 1973. 244 p.
(74) Stone, P.J. (and others): The General Inquirer: a computer approach to content analysis. Cambridge, Mass.: M.I.T., 1966. 651 p.
(75) Streatfield, D.: Moving towards the information user: some research and its implications. Soc. Sc. Inf. Stud. 3 (1983), p. 223–240.
(76) Tenopir, C.: Full text database retrieval performance. Online Rev. 9 (1985a) No. 2, p. 149–164.
(77) Tenopir, C. Searching Harvard Business Review online . . . Lessons in searching a full-text database. Online 9 (1985b) No. 2, p. 71–78.
(78) Trenner, L., Buxton, A.B.: Criteria for user-friendliness. In: 9th International Online Information Meeting; 3–5 December 1985; London. Oxford, Eng.: Learned Information, 1985. 486 p., p. 279–287.
(79) Wilks, Y.A.: Grammar, meaning and the machine analysis of language. London: Routledge, 1972, 198 p.
(80) Wilks, Y.A.: Relevance, points of view and speech acts: an artificial intelligence view. Memoranda in computer and cognitive science. New Mexico State University. 1985.
(81) Wormell, I.: Cognitive aspects in natural language and free-textsearching. Soc. Sc. Inf. Stud. 4 (1984), p. 131–141.

**Meeting on Concept Analysis and Methodological Problems of Psychology**

The meeting was organized by the "Forschungsgruppe Begriffsanalyse, Fachbereich Mathematik, Technische Hochschule Darmstadt" together with the Special Interest Group "Concept Analysis" of the German Society for Classification. It was the second of a series of conferences and it took place from Feb.13-14, 1987, with 40 participants. The first paper by M.BÖTTNER, Bonn, ("Begriffe als Prozeduren") provided an introduction into the newest developments in the area of procedural semantics with concepts being understood as procedures yielding a close relationship to cognition psychology. In his paper ("Kontexte und pragmatische Semantik") J.SCHÄFER, Darmstadt, discussed critically formal conceptual analysis from an intuitionistic and constructivistic point of view. B.GANTER, Darmstadt, showed (in "Abhängigkeit mehrwertiger Merkmale") how implications and dependencies between attributes may be studied with the methods of formal concept analysis. An adequate analysis of qualitative and quantitative data by trees was given by J.BANDELT, Hamburg, in his lecture ("Warm sind Baumhierarchien zur Repräsentation numerischer oder qualitativer Daten angemessen?"). J.ZELGER, Innsbruck, explained his method of philosophical context analysis (in "Eine Methode zur Entwicklung und Beurteilung von Forschungsprojekten") and showed its applicability to project work in groups. K.E.WOLFF, Darmstadt, demonstrated (in "Begriffsanalytische GRID-Auswertung") how formal concept analysis can be used in studying Grid-data. R.WILLE, Darmstadt, clarified (in "Formale Begriffsanalyse von Paarvergleichen") the combination of mathematical and interpretational questions by conceptually analyzing paired comparisons. Finally, an extensive discussion took place on the topic of the meeting, with questions and requests (by B.SEILER, Dannstadt, S.HÖPPE-GRAF, Heidelberg, A.CLAAR, Darmstadt, and S.STADLER, Basel) from the point of view of psychology. They pleaded unanimously for an adequate treatment of problems in the area of psychology by mathematical methods. It was encouraging that approaches to common work became visible.

In August 1987 the volume of the first meeting of Jan.1986 will appear under the title "Beiträge zur Begriffsanalyse" (Contributions to Concept Analysis), edited by B.Ganter, R.Wille and K.E.Wolff, and published by Bibliographisches Institut, Mannheim.

Rudolf Wille