JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE, 46 (3): 225-234, 1995

DOCUMENTARY ABSTRACTING: TOWARD A METHODOLOGICAL MODEL

MARIA PINTO MOLINA
Departamento de Documentación
18071 Universidad de Granada. España

ABSTRACT

In the general abstracting process (GAP), there are two types of data: textual, within a particularly framed trilogy (surface, deep, and rhetoric); and documentary (abstractor, means of production and user demands). For its development, the use of the following disciplines, among others, is proposed: linguistics (structural, transformational and textual), logic (formal and fuzzy) and psychology (cognitive). The model for that textual transformation is based on a system of combined strategies with four key stages: reading-understanding, selection, interpretation and synthesis.

A preliminary version of this work was presented at the 46th FID Conference on New Worlds in Information and Documentation, Madrid, 26-30 October 1992

1. INTRODUCTION. STATE OF THE ART

Writing abstracts is one of the documentary procedures that most facilitates the accumulation of knowledge in the face of the continuous outpouring of scientific literature. It is justified by the inescapable limitations of humans as a learners and containers of knowledge. The reduction of information is a natural consequence of the mind's entropic nature: Information is retained for very short periods of time, and only that considered important or significant is retained. The process of abstracting facilitates the acquisition of knowledge by going from specific to general, eliminating the incidental and conserving the essential.

As Winograd (1984) pointed out, abstracting competence entails the ability to explain the main points of a document in a brief manner. The general abstracting process (GAP), which refers to a chain of operations, implies the metamorphosis that textual documents must undergo from their surface and rhetoric structures, to the description of their deep (content) structure. It entails comprehensive, interpretative, selective and constructive processes, the purpose of which is the reconstruction of textual information as a new and representative document on a reduced scale. Abstracting is a complex operation. To obtain the abstract, the document text (which, as a whole, has a tree-shaped structure), must undergo "pruning": removing all the accessory foliage and maintaining the essential part, the "sap" or informative gist. The effort of producing an abstract from a text, or briefly expressing its "content", implies diving into the depths of the ideas pool, suprasensitive realities that can only be captured using the mind. The process is cyclical; it involves a journey from the concrete to the abstract, and back to the concrete, from an apparent world of words to an underlying world of concepts, intentions and feelings, and then a return to

the verbal surface. What we really aim to obtain is the metatext corresponding to an original prototext.

Cremmins (1982) and Rowley (1988), simplified this GAP and propose a pragmatic model based on global ideas. Cremmins (page 15) referred to it this way: "the mental acts are performed in four approximate stages: focusing on the basic features of the materials to be abstracted; identifying relevant information; extracting, organizing and reducing the relevant information into a coherent unit, usually one paragraph long; and refining the completed abstract through editing".

Rowley mentioned these five steps: read the document with a view to gaining an understanding of its content and an appreciation of its scope; make written notes of the main points made in the document; draft a rough abstract from notes recorded; check the draft abstract for punctuation, spelling, accuracy, omissions and conciseness; write the final abstract. But such strategies lack real scientific rigor. What is needed is a scientific model, an interpretative-selective model based on linguistic, logical and cognitive paradigms.

According to Boret and Peyrot (1969), the elaboration of any abstract will be governed by the following criteria: faithfulness to the original, which should be respected with regard to its content; precision, with the use of correct terms; explanatory clarity, by using the appropiate terminology; and entropy, giving the text the fullest meaning with the least number of words. As Maizell, Smith and Singer (1979, pages 80,92) emphasized "the good abstractor is fully aware that readers of his abstracts are likely to be busy people. Therefore, he strives for abstracts that are concise and succinct (but complete). He does not waste words. He avoids repetitive and meaningless expressions...” Likewise, for better readability of the abstract, the good abstractor "puts the essence first, writes concisely, writes clearly and understandably, (and) provides the full reference citation". In short, "a well-written abstract must convey both the significant content and the character of the original work" (Borko & Bernier, 1975, page 69).

In recent years there has been a strong trend toward automating the abstracting operations, but so far the results have not been completely satisfactory. As Paice states (1990, pages 175,182), "it is easy to appreciate that computer produced extracts suffer from two main problems: lack of cohesion and lack of balance"; he concluded that "it seems that progress with automatic abstract generation must depend on the existence of a satisfactory theory of text structure". Our purpose here is to study the process of abstracting from a manual and speculative point of view, because we are conscious that "radical analysis of the meaning of the text is not attempted". Only by the study of the text itself can we clarify the abstracting process. Such clarification is an essential first step toward successful automation of the process.

VARIABLES

The large number of variables that affect the GAP can be classified as textual, directly deriving from the textual reality, among which the base document or text is included; and documentary, generated by the documentary surroundings in which abstracting is conducted, such as abstractor, means of production and user demands. The word "text", as a scientific term that does not correspond to its everyday usage, refers to a "group of linked linguistic units in a total

conglomerate of communicative intention" (Petöfi and Garcia Berrio, 1978, p.56). Its defining principles are those of cohesion and coherence among its constituents. A text has a texture, and this is what distinguishes it from something that is not a text. The texture is provided by the cohesive relation (Halliday & Hassan, 1976). The text can be considered a complex cognitive and social phenomenon. The complex text structure follows a tree-shaped outline (the tree simile seems to be an extremely appropriate instrument to create a clarifying image of some of the basic concepts of textual linguistics), and implies the superimposition and interrelationship of the two basic structures (*surface* and *deep*) and a third one (*rhetoric/schematic*) that is complementary. The gist of textual being is based on this structural trilogy. The surface structure (microstructure) corresponds to the physical reality of the text, and its basic meaningful symbols, words, have the capacity to project themselves on our senses, allowing the perceptive process that culminates with understanding. The deep structure (macrostructure) is conceived as a hierarchical and coherent topic representative of the textual unit; it involves minimal structure of syntactic-semantic text representation.

The *rhetoric/schematic structure* (superstructure), a type of conventional production scheme to which text is adapted, can be considered as a transition between surface and deep structures. Van Dijk (1978) emphasizes that such "superstructures" are not arbitrary; rather, they reflect specific cognitive, pragmatic or social functions in textual communication. The study of these rhetoric structures is the only way to reach the textual taxonomy that is an essential prerequisite to the scientific study of text.

For the purpose of summarization, the distinction between narrative texts and expository texts is important. Narrative texts have is a "plot" and the task of summarization becomes one of determining what the "plot" is. Expository texts do not have "plots" per se and the events described may all be relevant. The task in summarization, then, becomes one of determining the appropriate level of detail (Rau, Jacobs & Zernik, 1989). There is a natural temptation to regard the rhetorical structure of a text in terms of its presentational structure- that is, its system of sentences, paragraphs, subsections and main sections. In actual fact, caution is necessary here, because the presentational structure is only a partial, and sometimes rather inaccurate, reflection of the true rhetorical structure (Paice, 1991).

Among the expository texts, scientific text has certain specific qualities, such as the *OMRC rhetoric structure*, a reflection of scientific activity itself, always concerned with the eternal sequence of research (objectives, methodology, results and conclusions), and the priority given to the implicit, already known, archive of information accumulated for centuries through documentary tradition. Because we usually presuppose a large quantity of information, presupposition is one of the most important factors to be considered in producing an abstract in science. Because of the enormous diversity in types of text and the lack of a true textual taxonomy, this discussion will be restricted to the scientific context, where the content is almost completely determined by the author (Kircz, 1991). It is worthwhile differentiating texts produced in the surroundings of natural sciences (NS) from those derived from social sciences (SS) and humanities (H), because there are differences in rhetoric among them, above all with regard to

methodology and the importance given to textual production as a channel of expression. Analysis of the rhetoric structure of scientific documents, based on OMRC, has confirmed the differences between a normalized and highly structured style of writing in the case of the NS, and the more idiosyncratic style of the authors in SS and H (Milas-Bracovic, 1987). Liddy (1991) claims that a prototypical empirical text has a discourse level (rhetoric) structure that has seven major components: subjects, results, purpose, conclusions, methodology, references and hypothesis.

An abstractor should also know the way in which textual content is revealed to the reader's comprehension (Armogathe, 1988). According to the linguist Jakobson (1971), language has six cardinal functions for some linguistic expressions, including text:: referential, emotive, conative, phatic, poetic, and metalingual.. Although none of these functions is found isolated in an enunciation of any length, the document profile characterization is achieved through the discovery of the more evident functions.

On the other hand, the nature and extension of the communicative value of linguistic units are substantially modified by contextual insertion. In fact, different types of context give, take away and change the meaning of messages. Context, in the general sense of the word, is a principle of incalculable linguistic value and provides the reality of expression, the Chomskyan performance, with that invaluable richness and variety. With regard to this context, it is worthwhile differentiating the following contextual strata: individual, documentary, social and cultural.

Among the possible qualities of abstractors, let us concentrate on two: memory and intellectual capacity. Memory represents the capacity to recognize and to remember. Two types are distinguishable: short term, based on the action of reverberation circuits, that are used to accumulate information with a "short life-span", with a surface structure; and long term, with greater possibilities for preserving information, the prime objectives of which are the semantic structures available during a much longer period of time. This is also called semantic or conceptual memory, and constitutes what some experts define as prior knowledge, or *base* knowledge, a notion that is linked to the theory of "schemes" ("frames"), conceptual structures referring to stereotypes. But the understanding of a text not only depends on the abstractor's capacity to store information in his memory, but also on the power of reasoning (inference included): the intelligence, or aptitude to create intellectual relationships, while it depends on the nervous system, is of a conceptual, as well as a sensory, nature.

INTERDISCIPLINARY CONTRIBUTIONS

The interdisciplinary nature of GAP is one of its most distinguishable characteristics. The role of linguistics in the consolidation of abstracting concepts and techniques is unquestionable because the processing of information is essentially conditioned by language, "a system which mediates, in a highly complex way, between the universe of meaning and the universe of sound" (Chafe, 1977, p.15), involving the most important form of known symbolic expressions. The linguistic factors (phonological, syntactic and semantic-pragmatic) affect respectively the

form, structure and meaning of texts. Although at first sight one might suppose the problem of textual analysis can be wholly solved by concentrating on meaning, it is undeniable that semantics is firmly associated with the phonological and syntactic factors, situated on the discursive communication "surface", which are an obligatory starting point. We must remember Gardin's (1973) claims that our only hope of understanding the intellectual operations implied in documentary content analysis is through the study of textual analysis in its different operative strata: word (paradigm), phrase (syntagm) and the text as such. See figure 1.

The major contributions of linguistics to the field of abstracting can be identified as follow:

1) structuralism, based on the work of Saussure (1916), conceives language as a structured system, an organized group of symbols (the union of a significant and a signifier), that offers the possibility of a functional analysis of language based on paradigm. The study of words, or paradigms, has been the obligatory starting point of semantics, and hence, of all processes of abstracting.

The greatest theoretical contribution has come from European structuralism, and from its theories on linguistic and lexical fields, which constitute the great revolution of modern semantics, and is based on the fact that articulation is the most general and deepest essential characteristic of any language, in accordance with the Saussurian idea of language as a system. The lexical field represents an articulated whole, a structure that reflects the eternal problem of synonymy, and above all that of antonymy. According to Weisgerber (1964, p. 71), “the greatest importance of the idea of field is that it has become the central methodological concept of research applied to linguistic content". Analysis by fields is the most effective means of determining content. In this same line of research on content analysis are the linguists Coseriu (1977), Greimas (1966) and Pottier (1974), who base their research on the Hjelmslev (1963) principle of linguistic economy (construction of an unlimited number of linguistic symbols by means of a limited number of nonsymbols, called figures), according to which a structural description cannot be made unless open classes can be reduced to closed classes. According to these principles, the number of semantic primitives in a language is limited; however, by linking up the primitives, a potentially infinite number of symbols may be constructed.

Of special importance is the use of anaphora, a technique for referring to an entity that has been introduced with more fully descriptive phrasing earlier in the text (Liddy, 1990). This is used quite naturally and frequently in both written and oral communication, to avoid excessive repetition of terms and to improve cohesiveness of a text, by means of: central pronouns, nominal demonstratives, relative pronouns, nominal substitutes, indefinite pronouns, pro-adjectives, pro-adverbials, subject references and definite articles.

The small number of words that constitute a large part of any text, are not the most significant. One must distinguish between function words, which express syntactic or operational relationships and which can be disregarded for certains purposes, and content or informative words. In spite of some progress in this area, vocabulary, the last linguistic layer immediately prior to the extralinguistic reality, is in need of much further investigation.

2) transformational (generative) grammar. Following the work of Chomsky (1986), on a fundamental system of linguistics based on the capacity (competence) of all language to

generate an unlimited number of phrases, language is now studied on two levels: One, situated in the deep structure, is patent. All text has this double structure, and complex transformation processes mediate between the structures. Transformational-generative linguistics give a much more definite focus for understanding a sentence. One can generally assume the deep structure generates the surface structure; between both of them lie the complicated processes of transformation. In any case, the component with generative capacity is the syntactic one, made up by the base, or group of rules that generate the deep structures, and the transformations, or rules that convert the deep structures into surface structures.

It is worth pointing out that the transformations are contextual rules because, in their formulation, there are certain restrictions or conditions that are imposed by the context. In any case, the meaning of a sentence depends on its syntactic structure and on the specific meanings of its elements. Thus making it possible to distinguish structural phrases, which cannot be either substituted or suppressed without destroying the text; the circumstantials *or* permutables, the suppression of which does not alter the deep structure of the text; and the stylistic elements, made up of lexical-rhetorical configurations. Structural phrases are more easily interchanged than word are and, in particular, are more easily translated from one language to another (Escarpit, 1981).

3) In the attempt to unify these two main trends, a new discipline emerged in the 1960´s, textology or text linguistics, which aims to widen the results obtained at the word or sentence level to the complete text, as a superior grammatical unit, and obligatory starting point in any task with regard to abstracting. To do this, the concept of sequence of sentences is introduced, as a group of textual units that possess cohesion and coherence. A particular degree of overall coherence is also demanded (Van Dijk, 1982). Text is defined from a functionalist point of view as the smallest unit possessing communicative autonomy. The principle of double structure, the root of transformational grammar, forms the basis of all textual theory. All linguistic entities, including text, reveal a double form of affirmation: a terminal-linear structure, the surface structure, and a deep structure. Between them is the rhetoric structure, the scheme on which the overall order of a text is based. These textual structures must be taken into account when designing a systematic research model for content analysis. Schemes based solely on "common sense" are not enough. In other words, the model must embrace rather complex theories that connect cognitive, linguistic, communicative, social and cultural parameters.

Language does not directly represent objects, but rather the concepts and objective proposals, being simply an intermediary of our communicative intentions. Moreover, it does not always adequately represent these concepts and objective proposals. It is an imprecise and ambiguous system in which the same linguistic symbol often represents different objective products (homonymy) and, vice versa, many symbols represent the same thing (synonymy). To control this linguistic reality, we traditionally rely on logic, one form of which, formal logic, carries out a large part of its work with symbols, that, unlike the linguistic ones, do mean exactly what logic needs them to mean. This form of working can be considered to be included in the domain of what we know as formalism; it consists of making an abstraction of the meaning of symbols and considering them exclusively as graphic units; that is what we understand by formalized

language, the extension of a method already known for centuries, the calculation method (to calculate well it is not necessary to know why we do it). Formal logic is a domain with enormous possibilities for application to content analysis of documents because it allows us to work logically with different units of meaning (functors and arguments) by assigning them to specific syntactic categories. The transformations derived from these logical operations with semantic units will guarantee the integrity of text as a content unit and, furthermore, will allow a correct and rigorous logical-semantic focusing of the analytical problem unknown in the linguistic-documentary field .

Our ways of viewing reality can be unduly restricted by the use of insufficiently flexible logics, because many aspects of human affairs, and of textual descriptions, are inherently imprecise; thus, the vagueness and ambiguity of natural language should not be viewed as imperfections or aberrations, but as central properties which need to be properly handled.

Hence the considerable growth in interest in fuzzy and imprecise logics, deriving from the ideas of L.A. Zadeh (1965), according to which the classes of objects encountered in the real physical world do not have precisely defined criteria of membership. Such imprecisely defined "classes" play an important role in human thinking, particularly in the domains of pattern recognition, communication of information, and abstraction. The concept of fuzzy set as a class with a continuum of grades of membership provides a convenient point of departure for the construction of a conceptual framework. One of the basic aims of fuzzy logics is to provide a computational framework for knowledge representation and inference in an environment of uncertainty and imprecision (Zadeh, 1993).

Fuzzy set theory attempts to generalize the traditional theory of sets by permitting partial membership (Bookstein, 1985). For each fuzzy set, a number between zero and one is assigned to each element of the universe under consideration, indicating the degree to which that element is in the set. Fuzzy set theory is a mathematical theory that provides methods and tools to allow one to grasp vague phenomena instrumentally. Therefore, it seems to be appropriate for use in the modeling of natural language semantics (Novák, 1993). The concept of fuzzy set may be seen as providing a new tool, more appropriate than that of classical set theory, for a program of summarization. It accommodates the inherent imprecision of the concepts that we actually use (Gaines, 1976).

Psychology, understood as the study of human conduct in its manifestations and in its structure, can also offer basic contributions to the study of textual content analysis, especially in its cognitive dimension (cognitive psychology), above all in understanding the complex mechanisms of knowledge acquisition and structuring. One can consider human knowledge can on a sensual level (through four elementary variables: sensation, perception, imagination and memory) and at an intellectual level (covering three realities: concept formation, judgment and reasoning).

OPERATIVE STAGES. STRATEGIES

In spite of the unity of the GAP, fragmentation is a coherent answer, not only to the sequential

reality of abstracting, but also, and above all, to certain pedagogical communication objectives. For abstracting, we propose an integrated model, taking into account contributions of several paradigms, of a pronounced cognitive-linguistic nature. In essence, it is based on three operative stages (figure 2):

Reading / understanding

Analysis: selection and interpretation

Synthesis / analytical description

Reading / Understanding

Reading, the only way of gaining access to documentary content, is a concurrent process, and not simply symmetrical to writing. It has an interactive nature, which depends as much on the text as on the reader, consisting of a series of coordinated procedures that include perceptual, linguistic and cognitive operations.

Among factors involved in reading, we can highlight (Pinto Molina,1992):

(1) The action of memory, which incessantly relates the unknown to the known.

(2) The participation of reason, and its complementary activities of induction and deduction, analysis and synthesis.

The action of reading is developed by means of the continuous and simultaneous application of two types of information processing, inverse and complementary, taken from cognitive psychology: ascendent, guided by data, inductive, bottom-up, in which reading is linear from the parts to the textual whole; and descendent, conceptually orientated, deductive, top-down, in which we move inversely by taking advantage of the base knowledge of the reader (Antonini & Pino, 1991). This is a double action (perception and understanding), and its strategies depend not only on reader and text, but also on the documentary objectives. These strategies may be cognitives, including automatic and unconscious interpretative behaviour, and metacognitive, or desautomated. Good reading, far from being a spontaneous act, must be organized and must follow a method. In abstracting, it is recommended that the analyst should first make a quick reading to recognize such fundamental characteristics of the document as form, class and structure of the information. In this first reading, although superficial, one is advised to take note of the relevant information. But a second reading will be necessary, performed carefully and actively, concentrating on the various headings of the document and on its key sections (purpose, methodology, results and conclusions) because these generally contain the deep and rhetoric structures of the document.

The traditional focus of studies of reading comprehension, based on the idea that reading occurs automatically when one knows how to structurally process the text, has given way to new theories that define textual comprehension as the process of creating a mental model that serves to interpret the facts described. This process depends basically on the inferences carried out by the reader in interaction with text because these inferences allow us to create a coherent text representation, connecting meanings of different and successive sentences. According to Schank (1979), inference is the nucleus of the understanding process and, for this reason, forms the centre of human communication. It is used to closely join the entries in a related whole. The inferences themselves are frequently the main point of the message. But the term

"inference" is not a precise one, because it is applied to a variety of different processes, numerous types being distinguished by the experts (Clark, 1977, Trabasso, 1981, Swinney & Osterhout, 1990). For our documentary purposes, the following types of inference should be considered: logical, because it is used to establish the causes, motivations and conditions that allow specific facts; evaluative, in which the analyst apply his/her beliefs to the situation described; of integration, carried out at the moment of understanding and based on the concepts and properties of hierarchical organization; and constructive, based on the abstractor's base knowledge. Anyway, the abstractor must go beyond the merely perceived, and must activate all kind of extralinguistic knowledge.

Therefore, we can at least deduce the following ideas: to understand is to integrate and interpret, it is to create meaning; the ascendent and descendent processes take part in understanding; depth of textual processing increases understanding; the perspective adopted by the reader greatly conditions understanding; and the process is not linear because conceptualized segments of discourse are constantly remodeled by the conceptualization of the following segments. The semantic is constantly translated into the conceptual.

Kintsch & Van Dijk (1985) base comprehension of linguistic enunciations on four principles: segmentation, from the continuous flow of signs of the language; categorization, a process which refers to the syntactic categories of the words, in the paradigmatic sense of language; combination, because said categories are juxtaposed, generating syntagmatic structures; and interpretation, because each unit of the language is assigned a specific meaning which is conventionally established. The understanding of information is based above all on this last activity, interpretation, which is only possible as a consequence of prior mental operations. The analyst applies these principles effectively, and must use what some authors call strategies, among which we can highlight presupposition, or the establishment of hypotheses on the text itself. The comprehension of sequences of sentences in a text must have a cyclical nature, corresponding to the cyclical principle of textual elaboration of information, which joins old (already known) and new pieces of information, by overlapping the different cycles.

Analysis

The analytical stage is the most difficult and controversial of the whole abstracting process because there is no rigorous and consistent methodology. In this analytical stage, we can differentiate two complementary activities: selection and interpretation.

Selection. Selection is a negative process; it consists of the elimination of meaning units (sentences and words) that are considered irrelevant for abstracting. Three groups of meaning units can be identified (figure 3): repeated, not very relevant and irrelevant.

Text must first be subjected to a process of contraction. Such a contraction, a frequently used method in the learning of languages and, particularly, expression techniques, consists of eliminating repeated (i.e., redundant) meaning elements. In theory, any text can be reduced to half its size, although contraction taken beyond a certain limit can result in the loss of textual sense. At this stage of contraction, the abstractor must take special care with anaphora, which must be both recognized and resolved.

Once contracted, text must be reduced. Reduction consist of the elimination of meaning

elements that are of little relevance. In the development of this selective work of reduction, because we are still dealing with the domain of language, linguistics can play a special role. The basic hypothesis or postulate of text linguistics is the well-known linguistic isomorphy principle, according to which the organization of our human communicative-verbal process is carried out by means of a quantitative expansion process which, however, scrupulously respects the structure of the elementary linguistic cell. The first linguistic reality of text is that of its construction, its production as synthesis. What abstracting aims to do is derived or secondary, in an inverse direction; the recovery of the analytical mechanisms from the basis of the finished textual product. Our abstracting objectives lead to a transcendental discovery: our methodological interest in uncovering the deep structure comes from the fact that both textual structures (surface and deep), although considerably different in size and extension, are cognitively equivalent.

The rules of transformation to be used in this reduction task, very similar to the rules of sentence generation in transformational grammar, obey the aforementioned principle of isomorphy, allowing us the necessary informative filtration. Said transformations will affect firstly the morpho-phonological strata, or surface structure; the lexical-syntactic; and the logical-semantic, giving way to the deep structure, the obligatory starting point in the interpretation process that follows (figure 4).

Once reduced, text must be condensed. Condensation consists of elimination of irrelevant meaning elements. At this point, the concept of relevance, closely bound to the documentary aims, becomes very important.

Interpretation. Once selected (contracted, reduced and condensed), text must be interpreted, assigning it a content. According to Brown and Yule (1983), the process of interpreting a speaker's / writer's intended meaning involves computing the communicative function, using general sociocultural knowledge and determining the inferences to be made. Needless to say, this is the most subjective moment of GAP because, apart from the objectivity of the textual content, certain extratextual factors now come into play, among them the base knowledge of the abstractor, the *context*, understood in the widest sense of the word, and the abstracting objectives.

Charaudeau (1983) explains that the success of the linguistic act depends on the symmetry or asymmetry between the process of production and the process of interpretation, involving two individuals with different competencies (to communicate and to interpret). This implies a semiolinguistic competence in which we can differentiate the linguistic component, made up of the different orders of organization of the language in its different conceptual devices; the situational component, made up of the sociolinguistic situations representative of social practices; and the discursive component, derived from the combination of the different conceptual devices, giving rise to multiple discursive effects, in accordance with the communicative intentions.

As Beghtol (1986) affirms, any text has a relatively permanent aboutness, but a variable number of meanings in accordance with the particular use that the person can make of said aboutness at any given time.

Synthesis

Once interpreted, text must be described. The synthesis, or analytical description, consists of the description of content derived from the analysis. This is the most delicate and difficult step; while the previous analytical activities can be subjected to more or less rigorous techniques, it is practically impossible to establish techniques of synthesis that are valid for all types of document and abstractor (Pinto Molina, 1993). This phase, according to some experts, is the authentic abstracting phase and is directly related to a specific property of natural language, namely elasticity of discourse. It is concerned with expanding the content structure obtained during the analytical (selective-interpretative) process, although this expansion must remain in the first stages of surface description because abstracts demand brevity. In this way, and after a repeated application of these mechanisms, we get the abstract, an autonomous secondary document, a brief and grammatically complete text which includes the original content from a documentary point of view. Its message has its own significance and importance. According to the American National Standards Institute (1979) and abstract of fewer than 250 words will be adequate for most papers and portions of monographs. For notes and short comunications, fewer than 100 words should suffice. For long documents, such as reports and theses, and abstract generally should not exceed 500 words.

TOWARD AN INTERPRETATIVE-SELECTIVE MODEL FOR ABSTRACTING

Most abstracting services offer guidance to abstractors that are oriented toward the product, the abstract, but pay little attention to the process of abstracting. The Chemical Abstracts Service (1989) guidelines offer only general content considerations to be kept in mind: the abstract must be representative of the technical content of the document; the abstracts should satisfy any questions posed or promises made in the title; the abstracts must be self-contained (understandable apart from, and without reference to, the original document).

Close to our position, Endres-Niggemeyer (1989) talks about basic rules for producing abstracts, differentiating the analytical ones (reducing and condensing) and those of synthesis (clarifying, reorganizing and stylizing): reduction is the elimination of the least essential components of the explicit meaning of the text; *condensation* is contraction without loss of explicit information.

Van Dijk and Kintsch (1978) define macrorules as the instruments that make possible the union between surface and deep structures. On the cognitive level, they are operations which tend to reduce the semantic information and are applied to the series of propositions that make up the text to obtain its general macrostructure. There are four main, or basic ones: omission, selection, generalization and integration. By means of the macrorule of selection one excludes propositions that are conditions, an integral part, presuppositions or consequences of the nonomitted proposition. In generalizing, the essential components of a concept are omitted by substituting one proposition with another, new one. A series of concepts is replaced by a superconcept that defines the overall group. In this way, what we normally think of as abstraction occurs. The fourth rule, construct or integrate, replaces the information with new information, in accordance with the principle of semantic implication.

Brown & Day (1983), following the Van Dijk & Kintsch (1978)model, propose a teaching methodology for abstracting based on five rules: removal of insignificant information, deletion of redundant information, concept superarrangement, thematic sentence selection (if possible, representing text), and abstract construction.

Having reached this point, and because of the extratextual factors involved in GAP, it is important to emphasize that the deep structure sought is subjectively variable, depending on the knowledge base of the abstractor and on documentary demands. This is a significant limitation when using the macrorules; in spite of having a general nature as principles of global organization and reduction of the information, they can be applied in different ways for different types of text in different pragmatic contexts (Van Dijk, 1978). Consequently, the systematic formation of an abstracting theory will depend not only on the extensive study of the text, but also on the documentary context and the aforementioned base of knowledge.

It is difficult to demonstrate that these semantic macrostructures, the condensed expression of the prior meaning of a textual singularity, respond to rules of a normative nature because, for any type of text that is sufficiently complex, there are as many possible macrostructures as interpretative acts.

As can be inferred in this brief exposition, the problem of abstracting is deep and complex. Deep because we try to establish the relationship between two worlds (sensitive and intellectual) that are radically different, although strongly linked, and the transition betwen the two is very difficult to achieve. Complex because the documentary demands, directly conditioned by user needs and abstractor qualities (strongly linked to his or her knowledge base) add new ingredients to the problem.

The abstractor needs a set of techniques rather than a theoretically predetermined system. In brief, the four key steps of the GAP (figure 5), can be summarized as :

(1) Reading-understanding: The first and essential step, text reading- comprehension, is a basic and complex activity, the common territory of several scientific disciplines (linguistiscs, logic, cognitive psychology).

This stage, considered as a kind of first analysis or preanalysis, is an interactive process between text and abstractor (a match text-abstractor), strongly conditioned by the reader's base knowledge, and a minimum of both scientific and documentary knowledge is needed. Reading concludes with comprehension, that is to say, textual meaning interpretation. This first (general-neutral) interpretation is the starting point for any analytical process.

(2) Selection: Selection is a process of purposeful elimination.

Developed by means of contraction, reduction and condensation strategies, the aim of selection is to retain only the relevant information. At this point the concept of relevance becomes extremely important.

(3) Interpretation: Having performed the selection step, the abstractor has to make a second (intended-selective) interpretation, depending basically on documentary aims. The two main sets of tools for interpreting are inverse and complementary: deduction and induction, reasoning and inference. Reasoning involves security; inference involves

probability. These basic activities may be improved when applied in a fuzzy manner.Thus we arrive at an interpreted meaning corresponding to an original one (figure 6).

(4) Synthesis /Analytical Description: Any kind of synthesis to be done must be entropic, coherent and balanced, retaining the schematic (rhetoric) structure of the document. When synthesizing, the abstractor must take into account the prefixed analytical description level, according to the desired type of abstract.

CONCLUSIONS

As Hutchins (1987) states, there are strong reasons for more thorough studies of the processes of summarization, involving information scientists, linguists, researchers in artificial intelligence, and many others. The aim should initially be not so much "automation" but basic understanding. Summarizing is essential to both text understanding and to text production, and it is crucial to information and knowledge organization; but we are ignorant about almost every aspect of the processes involved.

In conclusion, there is still much to be done in order to define an operational model of abstracting having the desired accuracy and reliability. The abstracting problem has not been resolved. This exposition merely approaches the problem, outlining the first steps towards an empiric model that would provide Information Science with a wide and practical explanation of document representation as far as content is concerned. The progress experienced in this domain will be directly linked to the two poles that basically condition it: object and subject, text and man. First, we should recognize its dependence on the budding science of text, the contributions of which must be the compulsory starting point for any proposed abstracting model. The more that is known about text as documentary unit, the greater the possibilities when abstracting it. Cognitive sciences (including philosophy, psychology, artificial intelligence and anthropology) and their most recent contributions will influence the strategies to be followed. The more that is known about the cognitive processes that affect human species, the easier will be the work of abstractor as a specific agent.

ACKNOWLEDGEMENT

I am very grateful to the anonymous referees for their patience in reading my imperfect English and for their constructive comments on an earlier draft. Special mention to Professor Wilfrid Lancaster, of the University of Illinois, for his help on content and style.

Granada, Mars 1994

REFERENCES

American National Standards Institute (1979). ANSI Z39.14-1979: American National Standard for Writing Abstracts. 1-15.

Antonini, M.M., Pino, J.A. (1991). Modelos del proceso de lectura: descripción, evaluación e implicaciones pedagógicas. In Puente, A. (dir.) *Comprensión de la lectura y acción docente.* Madrid: Fundación Germán Sánchez Ruipérez, 137-160.

Armogathe, D. (1988). *La synthèse de documents*.(Documents synthesis) Paris: Bordas.

Beghtol, C. (1986). Bibliographic classification theory and text linguistics: aboutness analysis, intertextuality and the cognitive act of classifying documents. *Journal of Documentation*, 42, 2, 84-113.

Bookstein, A. (1985). Probability and Fuzzy-Set Applications to Information Retrieval. In M.E. Williams (Ed). *Annual Review of Information Science and Technology*.)v. 20, 117-151.) New York: Knowledge Industry Publications.

Boret, M., Peyrot, J. (1969). *Le resumé de texte*. (Text abstract). Paris: Chotard et Associés.

Borko, H., Bernier, C.L. (1975). *Abstracting concepts and methods*. New York: Academic.

Brown, A.L., Day, J.D. (1983). Macrorules for summarizing text: the development of expertise. *Journal of Verbal Learning and Verbal Behavior*, 22, 1-14.

Brown, G., Yule, G. (1983). *Discourse analysis*. Cambridge. England: Cambridge University.

Chafe, W. (1977). *Meaning and the structure of language*. Chicago: University of Chicago.

Charaudeau, P. (1983). *Langage et discours: elements de sémio-lingüistique.* (Language and discourse. Elements of semiolinguistics). Paris: Hachette Université.

Chemical Abstract Services (1989). *Abstracting guidelines for CAS document analysts*. Columbus, OH: Chemical Abstracts Service.

Chomsky, N. (1986). *Knowledge of language: its nature, origin and use*. New York: Praeger.

Clark, H.H. (1977). Bridging. In P.N. Johnson-Laird & P.C., Wason (eds.). *Thinking*. *Reading in cognitive science* (pp. 411-420). Cambridge. England: Cambridge University.

Coseriu, E. (1977). *Principios de semántica estructural*. (Introduction to structusemantics). Madrid: Gredos.

Cremmins, E.T. (1982). *The art of abstracting*. Philadelphia: ISI Press.

Endres Niggemeyer, B. (1989). Content analysis -a special case of text comprehension-. In S.Koskiala, R. Launo (Eds.) *Proceedings: Information, Knowledge, Evolution, (pp.103-112).* Amsterdam: Elsevier.

Escarpit, R. (1981). *Teoría general de la información y de la comunicación*. (General Theory of information and communications). Barcelona: Icaria.

Gaines, B.R. (1976). Foundations of fuzzy reasoning. *Int. J. Man-Machine Studies,* 8, 623-668.

Gardin. J.C. (1973) Document analysis and linguistic theory. Journal of Documentation, 29, 137

168.

Greimas, A.J. (1966). *Semantique structurale*. (Structural semantics) Paris: Librarie Larousse.

Halliday, M.A.K., Hassan, R. (1976). *Cohesion in English.* London: Longman.

Hjelmslev, L. (1963). *Prolegomena to a theory of language.* Madison: University of Wisconsin.

Hutchins, J. (1987). Summarization: some problems and methods. In K.P. Jones (Ed).*Informatics 9:* Meaning: the frontier of informatics. (pp. 151-173). London:Aslib.

Jakobson, R. (1971). Language in relation to other communications systems. In *Selected writings* (V.II: Word and language. P.703). The Hague: Mouton,

Kintsch, W., Van Dijk, T.A. (1985). Toward a Model of Text Comprehension and Production. *Psychological Review*, 85, 5, 363-394.

Kircz, J.G. (1991) Rhetorical structure of scientific articles: the case for argumentational analysis in information retrieval. *Journal of Documentation*, 47, 4, 354-372.

Liddy, E. (1990). Anaphora in natural language processing and information retrieval. *Information Processing & Management*, 26, 1, 39-52.

Liddy, E. (1991). The discourse-level structure of empirical abstracts: an exploratory study. *Information Processing & Management*, 27, 1, 55-81.

Maizell, R.E., Smith, J.F., Singer, T.E.R. (1979). *Abstracting scientific and technical literature: An introductory guide and text for scientists, abstractors, and management.* Huntington, N.Y: Robert E. Krieger.

Milás-Bracovic, M. (1987). The structure of scientific papers and their author abstracts. *Informatologia Yugoslavica*, 19, 1-2, 51-67.

Novák, V (1993). Fuzzy sets in natural language processing. In R. Yager, L.A. Zadeh, (eds). *An introduction to fuzzy logic applications in intelligent systems (pp.185-200)*. Boston: Kluwer Academic.

Paice, C.D. (1990). Constructing literature abstracts by computer: techniques and prospects. *Information Processing & Management*, 26, 1, 171-186.

Paice, C.D. (1991). The rhetorical structure of expository text. In K.P. Jones (ed.) *Informatics* 11: The structuring of information (pp. 1-25). London: Aslib

Petöfi, J., & Garcia Berrio, A. (1978). *Lingüística del texto y crítica literaria*: (Text linguistics and literary criticism). Madrid: Alberto Corazón.

Pinto Molina, M. (1993). *Documentary analysis: basis and procedures.* Spanish edition: *Análisis Documental: fundamentos y procedimientos* (Documentary abstract: uotlines and methods). Madrid: Eudema

Pottier, B. (1974). *Linguistique génerale: theorie et description*. (General linguistics: theory and description). Paris: Klincksieck.

Rau, L.F., Jacobs, P.S., Zernik, U. (1989). Information extraction and text summarization using linguistic knowledge acquisition. *Information Processing & Management*, 25, 4, 419-428.

Rowley, J.E. (1988). *Abstracting and indexing*. London: Clive Bingley.

Saussure, F. (1916). *Cours de linguistique generale*. (General linguistics course) Genève: Bally & Sechehaye & Riedlinger.

Schank, R.C. (1979). El papel de la memoria en e procesamiento del lenguaje. (The memory role in language processing). (in Spanish) In N.Cofer, (ed.). *Estructura de la memoria humana.* (Human memory framework). Barcelona: Omega.

Swinney, D.A., Osterhout, L. (1990). Inference generation during auditory language comprehension. *The Psychology of Learning and Motivation, 25, 17-33.*

Trabasso, T. (1981). On the making of inferences during reading and their assessment. In J.T Guthrie (ed.). *Comprehension and Teaching: research and reviews.* Newmark: International Reading Association.

Van Dijk, T.A. (1978). *La ciencia del texto*. (Text science) Barcelona: Paidós.

Van Dijk, T.A. (1982). *Text and Context. Explorations in the Semantics and Pragmatics of Discourse*. London: Longman.

Van Dijk, T.A., Kintsch, W. (1978). Cognitive Psychology and discourse: Recalling and Summarizing stories. In W. Dressler (ed.). *Current trends in Text in Linguistics.* (pp. 61-86)Berlin: De Gruyter.

Weisgerber, L. (1964). Das Menschheitsgesetz der Sprache als Grundlage der Sprachwissenschaft. (Rules of languaje in use as basis of linguistics sicence). Heidelberg, p.71.

Winograd, P.N. (1984). Strategic difficulties in summarizing texts. *Reading Research Quarterly*, 19, 404-425.

Zadeh, L.A. (1965). Fuzzy sets. *Information and Control*, 8, 338-353.

Zadeh, L.A. (1993). Knowledge representation in fuzzy logic. In R. Yager, & L.A. Zadeh, (eds). *An introduction to fuzzy logic applications in intelligent systems*. (pp. 1-25). Boston: Kluwer Academic.