The SACAO Project: using computation toward textual data analysis in the social sciences[1]

Jules Duchastel, Luc Dupuy, Louis-Claude Paquin,
Jacques Beauchemin & François Daoust
Centre d'Analyse de Textes par Ordinateur[2]
Université du Québec à Montréal

1. The Project

The aim of SACAO (which stands for Système d'Analyse de Contenu Assistée par Ordinateur)[3] is to integrate systematically new or existing computer-aided reading procedures of textual data. It offers users various units of text description, exploration and analysis within a relatively integrated software environment, yet also lets them set procedures according to their own hypotheses. These procedures involve a minimum of theoretical preconstruct, thus allowing for maximal interaction between their application and the analysis of the text. Integration is ensured by establishing computer links between files sharing data structures. This friendly environment answers the many different needs of users faced with the intricacies of textual data analysis.

1.1 The problem at hand

Recent developments in computer science and natural language processing (NLP), the latter field which has yet to be clearly defined, not only concern the research community in the social sciences, but also the larger audience of users of the written word (documentalists, managers, decision makers, etc.). Personal computers have entered the halls of academe as well as they have corporate organizations, creating new work habits and generating growing amounts of textual information on magnetic support. This information can be found in data banks or in directories that still remain to be used to their full extent.

At this turn of events, users' expectations have risen with regard to the improvement of help procedures in writing and reading. In the production and management of texts, for example, they expect far more than just word processing. Some systems, already operational or in the first stages of development, assist users with their writing (through a lexical support of dictionaries, connectives, terminology, synonymy, etc.), revision (spelling aids, style sheets, etc.), and annotation (automated abstracting, indexing, thesaurus construction, etc.). Furthermore, problems encountered in accessing and enhancing textual data banks have also risen hope in reading assistance systems.

In all, these systems revolve around the description of the morphological, syntactic, semantic, logical or pragmatic make-up of texts, the search for relevant information or meaning within these, and the analysis of the data thus obtained.

On the one hand, we have computer uses of language and a growing amount of available textual data, and on the other varied reading and writing assistance procedures. But there is little methodology to support the integrated use of these procedures along a defined protocol. They are but only partial, not quite standard and often accessed with difficulty. Their use, if and when possible, is not strategic for lack of models designed to assist users.

1.2 The context

From the start, studies in natural language processing[4] centered on two lines of research: adapting linguistic and logical models to a computer environment, and perfecting "language engineering" techniques. Coulon and Kayser[5] placed each in its proper perspective: philosophy for one, which strives to further our knowledge of language, and ergonomics for the other, which is preoccupied with the production and use of tools. In the first instance, the plan is to program a computer to understand linguistic phenomena, in the second it is to provide the tools needed at each stage to facilitate this operation.

While the history of the field is permeated by both the philosophical and ergonomical perspectives, it has also witnessed a series of different theoretical approaches each of which has prevailed at a given time, then added itself onto the others and continued to develop to this day. The first years, from 1945 to 1955, are marked by a word frequency count approach. Syntax then predominates from 1955 to 1970, but as early as 1963 researchers are hard at work programming logical, semantic models. Since 1974, the main concern has been the representation and organization of knowledge according to cognitive models.

The reader will observe that these periods all refer to the classical stages in the understanding of language phenomena. Several types of research are found in philosophy and ergonomics. In philosophy we will point to the important growth of lexical approaches, the development of parsing techniques applied to restricted languages (LL(n) and LL(r) grammars), to formal syntactic models such as chain grammars, transformational grammars or even semantic grammars (case grammar, word function grammar, etc.). In ergonomics, software engineering has contributed among other things to the development of word processing, lexicon management, syntactic and semantic parsers (ATN), determining parsers, transformational grammars, definite clause grammars and finally, inference models. This is not meant to be an inventory but merely an indication of the wealth of fundamental and applied research found at each level.

Research has allowed for great strides in the field, but it has also pinpointed several problems. That one or another approach has prevailed at different times underlines the many lost hopes of having found the one approach to the processed understanding of language. Developments brought about by different fields and schools have also allowed for significant advances, but contradictions between theoretical approaches and the denseness of certain models have hindered the integration of the knowledge obtained. The relative short life of some projects betrays the existence of theoretical deadlocks. The problematic projection of theoretical hypotheses on practical applications has stressed the incompleteness of systems. Through this elaborate progression, however, the limits of linguistic, conceptual or interdisciplinary contributions that came to light allowed scholars to reevaluate the difficulties encountered in the understanding of linguistic and discursive phenomena. Some problems appeared more urgent than others, e.g. the contextualization of discursive phenomena, the representation of knowledge, the need to incorporate a considerate amount of extralinguistic data into NLP models, and the input of said natural logic.

2. The Approach Favored

We must first state that we narrowed our study to written language (including speech transcription), as opposed to spoken language, and to reading, not writing, aids. The approach then favored by SACAO centers on the following lines. First, it offers a pragmatic contribution to textual data enhancement, rather than a comprehensive approach to linguistic phenomena. Second, it stresses the analysis of discourse morphology instead of adopting a syntactic or semantic approach per se.

Regarding the first line of study, SACAO aims foremost at using operational modules on great bodies of texts. We have therefore chosen a pragmatic rather than fundamental approach or, to paraphrase Coulon and Kayser, the ergonomical over the philosophical perspective. The fundamental process is concerned with increasing knowledge first, and only afterwards will it seek solid and widespread applications to data from the "real world". The pragmatic process, on the contrary, takes interest in developing tools and applications that already enable us to extend our reading ability: quick and systematic access to the contents of great bodies of texts, precision and regularity in reading, production of new information, compared to the traditional forms of reading, introduction of validating measures and processes, etc. These have a practical value for one interested in the knowledge in the texts.

Even if we see fundamental and applied research as inseparable from one another, it is quite certain however that our aim to increase the potential for text content analysis falls within a pragmatic approach. But then again, there can be no application without some theoretical foundation bringing into play language, discourse and knowledge. Of course, practical choices made within SACAO cannot overcome this fact. We must at the very least question here the epistemological impact of our approach before coming back to the theoretical tenets that guide our project.

It would be improper in today's context to weave too tightly the fundamental process on the one hand with the "automated systems" applied to micro-worlds, and the pragmatic process on the other hand with the "assisted systems" applied to macro-worlds. Some studies in artificial intelligence, however, since they sought the isomorphic simulation of real phenomena, did focus on the automatic nature of procedures and worked toward system completeness. In its methodology, SACAO renounces the epistemological premises at the base of this approach. Automation is sought only on a pragmatic basis and has not become a prime goal. We favor instead an approach that is more hybrid and links automated and assisted procedures. Also, we substitute the concept of maximal tool integration to the purpose of system completeness. This approach is not only practical in that it is motivated by the necessity for a broad coverage of the real world, it also answers a wide definition of the problem of understanding linguistic and discursive phenomena and is based on a belief in the creative process that engages the user during the analytic process. Automated systems, however powerful they may be, are but a black box to users. SACAO offers an interactive method wherein researchers can invest their hypotheses and construct their analyses with the use of performing tools.

The epistemological stance of SACAO is thus empirically constructivist. In short, this approach conceives knowledge of linguistic phenomena as the product of a non-univocal object construction process. It implies the coexistence of many construction processes, complementary (e.g. the multiplication of levels of analysis) though potentially contradictory[6] (e.g. the coexistence of non-exclusively compatible approaches), and the need for interplay between model constitution and its empirical validation. Clearly, this process emphasizes the inductive method and the interactive nature of the system. We avoid, for example, projecting the model onto the data, and in a somewhat determinist fashion, projecting preconstructed theoretical models onto reality. We do lean, however, toward adding successive descriptions of the text in between our explorations of tentative results.

Let us return now to the theoretical aspects of SACAO. Two arguments prompt us to further explain our theoretical premises. One is that a tool production or selection process must necessarily find its coherence within a theoretical framework. The other is that, regarding the immediate interests of researchers involved in the SACAO project, a more theoretical course must regulate and bring into focus the developments later adopted. The second step in our approach refers to a theoretical presumption in favor of morphological discourse analysis.

One first theoretical choice, then, firmly sets SACAO in content analysis rather than linguistic description. Not that these two options are antagonistic, but giving priority to the grasping of meaning defines our workload accordingly to aim for the knowledge in the text. The stratification of levels involved in sociolinguistic phenomena (morpho-lexical, syntactic, semantic, logical and pragmatic) does not simply account for the many dimensions of language and discourse; it also offers a suitable step-by-step guide for reasearch. By methodological choice, both general and computational linguistics have often stressed the prime importance of the linguistic functioning of language and discourse. SACAO acknowledges the various levels of description as the result of the differential cutting and construction of the object; it does not consider them as organized stages in a compulsory process leading from lexical and syntactic description to total comprehension of natural languages.

By opting for morphological discourse analysis[7], the focus shifts from language to discourse. Linguistic description of a text will then serve as a support for the more complex analysis of its semiotic systems. We surmise that the text is a diversely structured space that unfolds according to a multiple sequentialization process (e.g. the narrative point of view, the argumentative point of view, etc.) in which objects are outlined to form cores of meaning. Our interest herein lies in locating the modes of segmentation characteristic of textual organization, and the condensations of meaning that occur at certain specific points. To this end, we refer to a lexical apprehension of the text inclusive of terminology, and to a non-exhaustive morpho-syntactic description of its units. We give greater importance to two main categories: nouns and verbs. The noun category is linked to the semantic organization of the text: analyzing context proximity or dependency (determination, topic-comment, etc.) allows us to reconstruct networks of meaning. The verb category, by contrast, is linked to the action structure of the text: analyzing the characteristics and environment of verbs to reconstruct the articulation and argumentation of the text.

3. The Methodology

It could well be said that the previous remarks set a course of research or a work space rather than define a clear conceptual framework. SACAO aims for minimal theoretical preconstruct precisely because it provides in place of an analytic model an environment that supplies an array of diversified and barely constrained means of reading. That is why we speak of a methodology for the integrated and strategic use of tools in textual data analysis. Integrated use is authorized by the architecture of the system which offers the choice of applying one or several procedures of text description, exploration and analysis and making them interact with one another in a larger context. Strategic use consists in permitting the choice of modules, their modification according to specific hypotheses, and encouraging total structuration of the research process.

While it does embrace an utilitarian approach, the system does not point toward plain automated understanding of the text. It supplies aids to text reading and analysis, giving the user tools that were tested in their present state of development. Its purpose is not to provide a method independent of user research context that would guarantee results generated through blind application of procedures. Instead, SACAO supplies tools that manipulate data whose theoretical premises have already been identified. The tools can and will be used in accordance with well-defined research strategies.

The system does indeed favor maximum interaction between the needs of the users and the reading and analytic aids that are provided to them. Users must be able to test the validity of results generated by any and all procedures available to them in order to select one. They must also be able to organize in the process proper recourse to the varied means at their disposition. If possible, they must choose the parameters that will be activated in each procedure. Conceiving these procedures will thus leave room for a redefinition of parameters.

The implementation of the system must take into account the characteristics heretofore mentioned. The main steps are the following, but first the feasibility of the project is possible only to the extent that we can rely on readily available software modules intended for text analysis, and on our own expertise in the field. Those modules are: SATO (a textual database system for content analysis), FX (a programming language for linguistic and cognitive systems), D-expert (an environment for expert system generation),[8] and various programs conceived for linguistic description (LCMF,[9] GDSF,[10] ALSF[11]). These systems were all developed at the Centre D'ATO, either by or in collaboration with its members.

Our research is based either on applications that are already developed or in the process of being developed (cf. the above-mentioned programs), or it can itself give rise to new developments. In the first instance, the modules are tested on large bodies of texts. This allows for the optimization of procedures or further for the identification of operational sub-modules whose use in text analysis is most important, e.g. categorization, topic and argument description. In the second instance, we account for original developments that are now known to be necessary for the general economy of the system. The "Set Phrases" and "Semantic Functors" modules are two such examples of developments now in the works.

The philosophy underlying SACAO is one of integration of the various modules according to the creation of links within the same computer environment and the transferability of modules from one environment to the next. Each adaptation of these modules as well as each new development should be integrated and implemented within any environment. In a more realistic vein, however, the prime goal must be to achieve integration of all modules in a global environment, while some particular ones will be available on personal computers.

We systematically and regularly test the various modules of SACAO on large bodies of texts. We own a large data bank consisting of texts we have gathered from several research projects. For the most part, testing is done with data obtained from the public domain. Without restricting its use to other types of application, this means that utility programs (e.g. terminological idiom dictionaries, semantic dictionaries, etc.) are first enriched by data from the public domain. The environment may then seem more familiar to the discourse analyst than to the literary critic. It must be mentioned, finally, that testing has brought about the systematic writing of technical reference cards which record at length the various procedures and will later serve in establishing a user manual for SACAO.

4. The Architecture of the System

4.1 The goals

On the computational level, the goals of the SACAO project are the following:

1. Strengthening of the system by allowing greater integration between modules. Providing transfer from one type of computer to another in order to enable users to perform certain tasks in a familiar environment while giving them access to greater processing capacity.

2. Evaluating existing modules systematically in order to enrich them or extrapolate specific procedures more pertinent for use. Improving procedures of description, exploration and analysis encompassing greater complexity and coverage.

3. Encouraging access to the system by supplying a detailed and exhaustive documentation of procedures based on rigorous testing of controlled bodies of texts. The functional dimension of the architecture of SACAO is described hereafter. We must acknowledge however the many dimensions of the term architecture. The functional dimension, on which we focus our attention here, describes the features of the various modules comprising processing units. We will not touch upon the organic and algorithmic dimensions.

4.2 User-computer interplay

At the moment, the software modules are implemented in different computational environments. Such a variety of work environments could account for great difficulties in the use of the resources of SACAO. To prevent any inconvenience, we chose two ergonomical principles to offset these difficulties: transparency and transferability.

Transparency must be ensured in order to provide the user with an interplay comparatively independent of the hardware used. In general, decisions are made in interactive fashion from a choice offered in stratified menus. This management "by menus" encourages user-software dialogue which must be open to the context.

To this principle, we must add that of transferability. This principle specifies that all options of development must assure the transfer of knowledge contained in management modules and processing units. Transferability from one hardware installation to another (PC to VAX-VMS or UNIX, VAX to MacIntosh, etc.) can provide access to cooperative processing (e.g. developing a blueprint for analysis on PC and pursuing data processing on VAX), data transfers between the different processing units, etc.

4.3 Textual data management

Since we wished to make the tools and textual data gathered within SACAO accessible to the greatest number of users, we studied the problem of data management. Our goal was to structure public-oriented program banks. These contain the array of modules used in textual data processing, and the procedures called upon for batch processing. They also integrate the bodies of texts that researchers have wished to make public. The mechanisms herein assume the cumulative aspect of tool production for textual data analysis.

To archive utilities we must add another program for proper ASCII format conversion to all hardware installations. With this program, French-speaking users are able to maintain the integrity of their texts and proceed similarly with the analysis and processing of data in all hardware installations.

4.4 Textual data description

Any mode of investigation implies technical intervention on the data to be analyzed. The very notion of "data" necessarily calls for a process of unit construction and, therefore, a re-structuring intervention that transforms information units into units for analysis. The textual data description module is where the initial structuring of the data occurs. Within the context of the SACAO project, three levels of description are foreseen: lexical, morphological and syntagmatic. Generally speaking, these levels are independent from one another, but they can be combined so as to answer the specific needs of a research hypothesis.

At the lexical level, data description aims to structure the different aspects of the vocabulary (or lexicon) of a text. One may think specifically of lexicon structuring from a phrase dictionary or a specialized thesaurus. In both cases, the procedures involve listing all elements in a body of textual data. One must then add to the core vocabulary of French phrases that reflect the idiomatic features of a given linguistic community. Lexical forms are often found in clusters that function as words. In order to assist userss in listing these units, the textual data description module gives them the opportunity to recuperate all the different forms of phrases to be found in their data. It is thus possible, within the lexicon of a large body of texts, to index set phrases (prepositional, adverbial, etc.), idiomatic phrases to one or a family of speakers, technical phrases, institutional terms, onomastic phrases (proper nouns), etc.

At the morphological level, we must work toward clearly identifying the grammatical dimensions (lexical and grammatical morphemes). We have at the moment a processing unit for the part of speach characterization of contemporary French (BDL)[12]. This unit enables us to index the elements of a vocabulary or a lexicon by completing lexical forms with syntactic labels (for classifying nouns, verbs, adjectives, etc.). A second unit (LCMF) enables us to mark features pertinent to the lexical dimension of words (lexical or radical morpheme).

Finally, we also have at our disposal processing units for describing the syntagmatic dimensions of textual data. At one level, we can call upon two programs capable of producing automatically or semi-automatically a syntactic phrase description ("grammatically correct phrases") of contemporary written French. The first program, GDSF, heuristic by nature, is able to detect in any proposition the topic, the comment, various indications on verbal complements and many types of noun determination. The second program, ALSF, still in development, encompasses a greater linguistic scope. Conceived as a global environment to process French idioms, it foresees modules for syntactic information, analysis and interpretation. In its present state, it offers some processing units,e.g. noun phrase description.

At another level, there are some examples of text analysis programs that either first rely on a morpho-syntactic description of phrases in the text, or on its semantic organization. One such example can be found in SAADI,[13] which functions on the basis of noun grouping and clause structuring (concessive, restrictive, conclusive, etc.) and is able to describe the argumentation of a text. There are also grammars in semantic representation of various textual objects that were developed by some researchers. Should our interests lie in levels of text structure other than the morpho-syntactic (e.g. thematic analyses, phrase classification, etc.), we then have units at our disposal with which we can program algorithms of description to tailor our needs. For instance, FX permits the programmation of automated or assisted grammars.

4. 5 Textual data exploration

The exploration module performs complementary work to that of the description module. Once their data has been gathered, users should have at their disposal a choice of specific operations for data selection, reassembly and classification. There are processing units in the information retrieval module for dressing inventories or for grouping information by categories.

For those units that are structured along a linear (lexical) sequence, we are then able to obtain: frequency lexicons; concordances (or KWIC: Key Word in Context) drawn from key word searches or symbolic or numeric tags attached to these key words; co-occurrences (e.g. key words and a lexicon of words closely associated to these key words), etc. In order to detect these expressions, we rely on a string of operations that can determine the form and number of character chains to be used as parameters in the information retrieval process.

For those units, however, that are structured along morphological constraints, either clearly defined (syntactic configurations, tree diagram structures) or not (thematic units, axiological statements, etc.), the information retrieval module allows users (researcher or analyst) to detect data from their own patterns.

In addition to dressing inventories and classifications, the exploration module can also define and circumscribe parts of the analyzed corpus. Thus, a user working on a body of texts can very well apply the search procedures mentioned above on arbitrarily defined subsets. In other words, it is possible to generate a variety of subtexts from one corpus. Text generation here can be made to meet the norms of statistical treatment (sampling techniques) or to allow users to verify their hypotheses on a comparatively restricted subset before pursuing the process on the body of their texts.

4.6 Textual data analysis

At this stage, the textual data analysis module offers the following processes:

A) A lexical-statistical module drawing the following statistics for any given lexicon: mean, standard deviation, variance, minimum and maximum frequency, z score, and procentual distribution of group frequencies and word tokens.

B) Intertextual distance measures. These enable the user to compare, two by two, texts or subtexts in order to establish which lexical elements are "responsible" for surface gaps between them. Intertextual distance can be analyzed according to different frequency distributions and dispersion corresponding to various segmentations of the lexicon and counterbalanced by a reference lexicon identified by the user.

C) Readability indexes.[14] These are empirical measures that estimate the difficulty or ease in reading, understanding and memorizing a text or parts thereof. They are calculated from such parameters as word length, phrase length, etc.

5. The Operation of the SACAO Project

Let us come back briefly to the main conclusions that emerge from the preceding exposition before showing their impact on the definition of the SACAO project and the organization of its activities. From the onset, we recognized the need for textual data reading assistance. This need is felt by scholars and researchers from the many disciplines that rely on textual material as a source of knowledge, as well as in organizational settings where text is used to various ends. We opted for an ergonomical approach to the question, advocating the integrated use of diversified tools as a support for analysis. By giving priority to content analysis over a purely formal knowledge of language, we favored an interdisciplinary approach. This pragmatic point of view encourages a heuristic stance in the research process and, as for the tools at their disposal, the greatest autonomy for researchers. By calling upon automated and assisted procedures as well, this hybrid philosophy calls upon an active participation of the text analyst.

The means at our disposal were then adjusted according to these needs and this approach. The elaboration of this methodology for the integrated use of reading aid procedures meant the development of an environment which allowed the strategic management of these means. End-users must be able to choose the procedures they will keep, and choose also the parameters that will be activated within these. They must be able to link, in many diverse ways and in keeping with their own needs, the various procedures between them, thus structuring as a whole their research process. In order to meet this demand, the specifications of the system were written so as to encourage interplay between the user and the tools. They remain open to some parameter variation and comprise the greatest document support for researchers.

The architecture of SACAO was conceived according to this orientation. It defines various strata that correspond in a way to the very research process of its users. By giving them standardized methods of operation and management facilities, it also delimits three main fields of activity around the description, exploration and analysis of textual data.

The SACAO project was conceived and developed in a context that adequately reflects the concerns we have underlined so far. Inscribed somewhat diffusely at first within the activities of the Centre D'ATO, it has since been specified in a differentiating process from other fields of research in natural language processing. Apart from the necessary development of linguistic or cognitive description modules, the specific need for tools for text analysis was felt with great urgency. The SACAO team consisted of researchers whose disciplinary affiliation and fields of specialization were certainly different, but whose common goal was the analysis of texts. The feature of the team was that it corresponded to heterogeneous demands in terms of development. While some of our activities fit within the structure of university research, others were immediately associated with demands of developing systems geared toward corporate organizations.

This team, in which each member, moreover, independently pursued work within his field of specialization, then had to mount a common project that reflected the polymorphous aspect of the needs, the approach and the means advocated. It defined four fields of activity and installed the mechanisms needed for their realization. These fields were: computer development, adaptation and development of processing units, experimentation and documentation, and finally exchange and training. The means used consisted of a weekly seminar of exchange and planning, and a sharing of tasks according to each member's expertise. We will quickly illustrate what kind of activities pertained to each field.

Computer development refered to the computational aspect linked to the establishment and management of reading aid procedures. It can be a matter of maintaining software environments in the various installations, or adjusting interfaces and transferability. It pertained also to various computer developments linked to those of procedures: new structures of representation, new automations, etc. It also included the development of file management procedures.

The adaptation of processing units can be illustrated by an evaluation study we made of GDSF descriptions of the thematic and determination structure of texts in a body of political speeches[15]. On the basis of this validation, enriched by new developments, some subsets of procedures were used to establish a tree diagram description of sentences according to their thematic or determination hierarchy in the functional grammar tradition. The development of new program modules can be illustrated by the new procedures of phrase detection and thesaurization. This system used features of our programs in order to supply a new tool to users.

Experimentation refered to the systematic process of procedure validation on reference texts. This work let us vary the contexts of application and test the strength of systems when redefining parameters. Apart from validation, experimentation enabled us to produce technical reference cards intended to document the system and end-user reference cards intended for researchers.

Finally, exchange and training activities appeared most important to us. Since our project was interdisciplinary and given that it explored multiple research venues, we were forced to agree on a series of theoretical and methodological questions. The topics we studied pertained to problems in semantic categorization, strategies in discourse analysis, approaches in thematic analysis, the parsing theory, etc. Training was given through specialized courses in ATO.

In all, SACAO was not an insular project, but an open research program. It corresponded to precisely identified needs and provided an interdisciplinary work space. Even though it benefited immensely from fundamental research in computational linguistics and cognitive science, SACAO never lost its focus, which was computer-assisted text analysis.

6. Post-SACAO developments

Two projects followed from SACAO. The first project, ACTE (Workbench for knowledge engineering and textual data analysis), is under the supervision of researchers[16] from the Centre d'ATO and training officers in public organizations. It is commissioned by a consortium of Departments in the Government of Quebec. In large organizations such as those of the government, textual production - in the form of reports, guidelines, memos, etc. - is increasing at a rate that hinders its efficiency. "Text workers", researchers, managers and decision makers whose main activities are reading and analysing texts, are thus submerged by a mass of documents they must analyze according to specific objectives : accumulation of facts, events or knowledge, interpretation, stategy planning, decision making, etc. On the other hand, computer tools and methods for understanding texts have been developped and further improved in previous research projects, as SACAO. The main goal of ACTE is to give access to an integrated environment of cognitive and textual system to help solve those problems.

The project is specifically oriented toward the integration of primary textual sources in a process of knowledge engineering. Knowledge is extracted and formatted from the textual base to become either goals, facts, rules or inference mechanisms. The system is designed as an interactive environment for interfacing standard and textual databses, a textual parser and an expert system generator. The textual databases and parser is managed by SATO and, once the textual data has been properly translated into goals, facts or inference rules, the data is taken in charge by the D-expert environment for the generation of expert systems.

The second project, HERMES (Intelligently assisted document management), is still at the early stage of funding[17]. As is the case for the ACTE project, it attempts to solve the problem of processing large number and volume of documents in large organizations. Standard information processing systems are insufficient to take care of the actual content of texts. The general long-term objective of the HERMES project is to construct an integrated assisted analysis and computer management system for large electronic document databases. The data-processing infrastructure will draw on technologies already developed in previous projects (SACAO, ACTE), but will nevertheless have to be constructed almost from the beginning. In the first case, the textual analysis engine will be SATO with increased features, especially to increase its processing capacity in terms of gigabytes, and the textual analysis control engine will be ACTE. In the latter case, a specialized text database management system will have to be developed, as well as many modules for : linguistic description, code conversion, SGML markup, Indexation, calibration, etc.

The goal is therefore to develop the design of a computer platform with modules for intelligently assisted document management and for the analysis of these documents. The system is called "intelligently assisted" because it attempts in its operations on various types of documents (texts, images, sounds, animation, etc.) to help the user perform a large number of tasks generally associated with these documents, such as writing, standardization, analysis, classification, indexing, markup, description, knowledge extraction, retrieval and distribution. As was the case for SACAO, the system is modular in that it consists of a number of processing units performing specific operations, but whose inputs and outputs can always be communicated to other modules. There will be a high degree of interaction between these modules. The system will also remain under the ultimate control of a user who will decide on the strategies to use for navigation in the database and completion of the required tasks.

As one can see, the SACAO project has defined the general guidelines for software developments as well as a general methodology which can now be applied to much more important quantities and different types of documents. The different projects which followed, took over where the SACAO team had left off, namely the basis for a computer assisted system for content analysis.

BIBLIOGRAPHY

ALLC, "Méthodes quantitatives et informatiques dans l'étude des textes", Genève - Paris, Slatkine - Champion, 1986, 947 pages.

Allen, S. 1982. *Text processing: text analysis and generation: text typology and attribution.* Stockholm: Almqvist & Wiksell International.

Berwick, R. C. 1985. *The acquisition of syntactic knowledge*. Cambridge, Mass.: MIT Press.

Barret, E. 1989. *The Society of Text, Hypertext, Hypermedia, and the Social Construction of Information.* Cambdrige: The MIT Press.

Borel, M.-J., Grize, J.-B., and Miéville, D. 1983. *Essai de logique naturelle*. Berne: Éditions Peter Lang SA.

Coulon, D., and Kayser, D. 1986. "Informatique et langage naturel: Présentation générale des méthodes d'interprétation des textes écrits". *Technique et Science Informatiques*:103-126.

Cruse, D. A. 1986. *Lexical Semantics,* Cambridge: Cambridge University Press.

Danlos, L. 1987. *The linguistic basis of text generation,* Cambridge: Cambridge University Press.

Daoust, F. 1992. *SATO: Système d'Analyse de Textes par Ordinateur (version 3.6). Manuel de référence* Université du Québec à Montréal, Centre d'Analyse de Textes par Ordinateur.

Duchastel, J., Paquin, L.C., and Beauchemin, J. 1992 "Automated Syntactic Text Description Enhancement : Thematic Structure Analysis". *Computers and the Humanities* 26: 31-42.

Duchastel, J. , Paquin, L.C., Beauchemin, J. 1992 "Automated Syntactic Text Description Enhancement : Determination Analysis". *The New Medium, Research in Humanities Computing.* Oxford: Oxford University Press, (to be published).

Gross, Maurice, 1975. *Méthodes en syntaxe. Régime des constructions complétives.* Paris: Hermann.

Grosz, B. J., Jones, K. S., and Webber B. L. 1986. *Readings in Natural Language Processing.* California: Morgan Kaufmann Publishers Inc.

Halliday, M.A.K. 1985. *An introduction to functional grammar*. London: E. Arnold.

Krippendorff, K. 1980. *Content Analysis. An Introduction to its Methodology.* Beverly Hills: Sage Publications.

Marandin, J. M. 1988. "A propos de la notion de thème de discours. Éléments d'analyse dans le récit". *Langue Française.*

Melchuk, I. A., and Arbatchewsky-Jumarie, N. 1984. *Recherches lexico-sémantiques.* Montréal: Presses de l'Université de Montréal.

Paquin, L.-C. 1990. *D_EXPERT (Version 2.0)*, Université du Québec à Montréal, Centre d'Analyse de Textes par Ordinateur.

Plante, P., 1975. *Proposition d'algorithme pour le dépistage de relations de dépendance contextuelle dans un texte.* Montréal: Université du Québec à Montréal.

Rastier, F. 1991. *Sémantique et recherches cognitives*. Paris: Presses universitaires de France.

Rastier, F. 1989. *Sens et textualité*. Paris: Hachette.

Rastier, F. 1987. *Sémantique interprétative* Paris: Presses universitaires de France.

Sowa, J. F. 1984. *Conceptual Structures. Information Processing in Mind and Machine*, Reading, Mass.: Addison-Wesley.

[1] This text was translated from French by Dominique Michaud.

[2] The Centre d'Analyse de Texte par Ordinateur (Centre for Computer Assisted Textual Analysis) will be refered to, in the following pages, as the Centre d'ATO. The name has changed since june 1991 to Centre ATO-CI, Centre en Cognition et Information, refering to the cognitive (C) and information (I) dimesions of our actual research work.

[3] The project originated in 1986 but only started to be implemented in January 1988. From 1989 up to now, it has evolved towards two other main projects. One of them is ACTE (Atelier cognitif et textuel) which consists of the integration of a textual data base system for content analysis (SATO) and a expert system shell (D-Expert). The most recent one has just been submitted for research grants. The HERMES project is an intelligent system for the assisted information management and analysis of documentary bases. Those projects derive from SACAO and include

members of the original team as well as new researchers. The approach, the methodology and the main goals have been preserved in both projects as one can see in the brief outline made in section 6.

[4] Cf. on the subject the detailed analyses in Daniel Coulon & Daniel Kayser, "Informatique et langage naturel: présentation générale des méthodes d'interprétation des textes écrits", *Technique et science informatique*, 5, 2 (1986); and B.J. Grosz et al., *Readings in Natural Language Processing*, Morgan Kaufmann Publishers, 1986, 664 p.

[5] Daniel Coulon & Daniel Kayser, "Informatique et langage naturel", *op. cit.*

[6] This notion of eclectism and complementarity in approaches is also found in the epistemological thinking of quantum physics: cf. Fritjof Capra, *The Tao of Physics*, Boulder: Shambola, 1976; and *Le Temps du changement: science, société, nouvelle culture*, Paris: Éd. du Rocher, 1983; also Heinz Pagel, *L'Univers quantique*, Paris: Inter-éditions, 1985.

[7] We wish to acknowledge the important contribution of Alain Lecomte (GRAD, Grenoble, France) and Jean-Marie Marandin (INaLF, Paris, France) to the field of discourse analysis and specifically to the development of the hypotheses discussed here.

[8] SATO was conceived and developped by François Daoust of the Centre ATO-CI. FX was conceived and developped by Pierre Plante of the Centre ATO-CI. D-expert was conceived and developped by Louis-Claude Paquin. Daoust and Paquin are active members of the SACAO project.

[9] LCMF (which stands for Lemmatisation et Caractérisation Morphologique du Français), a program conceived by Lucie Dumas, enables the user to recognize the syntactic category of lexical forms in French an to gather automatically all inflected forms of a particular unit of representation.

[10] GDSF (which stands for Grammaire de Surface du Français), another program conceived by Pierre Plante, consists of a set of procedures designed to obtain surface structures in written French and programmed in Déredec (previously used as a language for programmation of automated grammars, akin to Augmented Transition Networks) .

[11] ALSF (which stands for Analyseur Lexico-Syntaxique du Français), conceived in collaboration with and under the supervision of Jean-Marie Marandin of INaLF, constructs the syntagmatic structures projected by the major grammatical categories of French: nouns, verbs, adjectives and prepositions. It also constructs the relations these categories entertain with each other in sequential units.

[12] BDL (Which stands for Banque de Données Lexicales), conceived by Luc Dupuy of the Centre ATO-CI, gives morpho-syntactic information out of context for about 358,820 words in written French.

[13] SAADI (which stands for Système d'Analyse Assistée des Interviews), conceived by Alain Lecomte and Catherine Péquegnat of the Université de Grenoble, takes into account question and answer sequences and is able to detect direct answers in the context of an interview.

[14] These indexes are discussed at length in François Richaudeau, Le Langage efficace, Paris: CEPL, 1973, 300 p.

[15] See DUCHASTEL, J., PAQUIN, L.C., BEAUCHEMIN, J., "Automated Syntactic Text Description Enhancement : Thematic Structure Analysis", *Computers and the Humanities*, no 26.1., 1992 and DUCHASTEL, J. , PAQUIN, L.C., BEAUCHEMIN, J., "Automated Syntactic Text Description Enhancement : Determination Analysis", *The New Medium*, *Research in Humanities Computing*, Oxford University Press, to be published.

[16] Those researchers are, for most part, the same as in the SACAO project.

[17] The project is under the supervision of Suzanne Bertrand-Gastaldy and Jean-Guy Meunier but involves most of the SACAO team as well as many other researchers. It also involves a partnership with the industrial and governmental sectors. It has been submitted to the Government of Quebec's SYNERGY program, which finances research and development intiatives in the area of technology transfer. The project, if realized, will require significant funding.