An Approach to Expertise Transfer : Computer-Assisted Text Analysis

LOUIS-CLAUDE PAQUIN & LUC DUPUY
Université du Québec à Montréal[1]

0. Abstract

This paper aims to contribute to the efforts made by researchers in the automated or assisted transfer of textual knowledge to the expert system. The main objectives of our research approach are constance, objectivity, reproductibility and independance of problematics exposed in the texts. This is why we developped a morpho-syntactic text analysis method to locates significant concepts in a large corpus. This method is inspired by discourse analysis and emphasizes a unified and "natural" perception of knowledge.

1. The Context

Expert systems allow us to tackle a series of problems for which conventional computer solutions are of little help, owing to the socio-cognitive features of expert knowledge. Data handled by expert systems often appear in a complex and incomplete fashion; data structuring is also subject to frequent review. The principle underlying these systems is one of separation between domain semantics knowledge and computer instructions that summon relevant knowledge units at the appropriate time.

Today, one might very well argue that expert systems have achieved a satisfactory level of sophistication and stability. We must however emphasize the enormous costs implied in those operations of locating and apprehending knowledge. This in fact constitutes a majorobstacle in the introduction of this particular computer technology in organizations.

This paper[2] aims to contribute to the efforts made by researchers in the automated or assisted transfer of human knowledge to the expert system. Many applications are inspired from cognitive psychology: some are based on the personal construct theory,[3] in which the expert system assesses objects in triads and then generates a distance matrix from which a conceptual hierarchy will be drawn; others are based on protocol analysis,[4] the verbal comment produced during problem resolution. Drawing on the neurophysiology of the nervous system, a neural network approach[5]

allowing a "literal" apprenticeship of the system by accumulation of stimuli is under development in laboratories. Another venue of research is morpho-syntactic text analysis,[6] which locates significant concepts in a large corpus. The approach we adopted is derived from the latter. It is inspired by discourse analysis, which emphasizes a unified and "natural" perception of knowledge. Our approach permits a constant, objective, reproducible and independent screening of problematics exposed in the texts.

The genesis of the present paper is twofold. In our capacity as knowledge and textual "engineers", we have been working on several pilot project commissioned by government agencies. Among others a knowledge based information system dovoted to environmental evaluation management of large projects.[7] The relevant knowledge, almost exclusively textual in nature, occupies a significant volume of space:[8] it consists of legislative texts, regulations, directives, decrees and correspondence. In addition, thirteen hour-and-a-half long interviews were conducted with researchers involved in the project and later transcribed into data. We are also involved in research and developement projects devoted to textual analysis by computer.[9] In all those intervention, the theoretical stance adopted stresses morphological discourse analysis (discussed in further detail later on) over a stringent syntactic or semantic representation of textual data.

This paper is focused mainly on ascertaining the theoretical premises of our approach to knowledge extraction from texts and determining the conditions of its feasibility on a large scale. Given the complexity of the task, we decided to proceed in two steps. The first step, presented here, consisted in revealing the cognitive structures of the field of expertise: the result being a dictionary of concepts, structured as frames, constituted by analyzing substantives and the word clusters associated with them.[10] The second step should supply us the information we need to write inference rules and order the operations to be carried out with these. This aspect of the research will rely on an analysis of verbs of action involving the structured objects isolated beforehand, and on a division of the corpus into uniform segments that will constitute as many subtexts describing each stage in the resolution process.

In the following sections, we first touch upon the capture of expertise (#1-2). We then specify our intervention with regard to the texts (#3-4). Following a definition of knowledge, we propose natural logic as a context of reference (#5-6); this discussion opens onto a presentation of morphological discourse analysis, with special attention given to the objects of schematization (#7-8). These theoretical precisions lead us to describe the details of our mode of analysis (#9). Finally, we define each of the three stages in composing a dictionary of valued objects (#10-12).

2. Reconstituting Expertise

Unless he is a teacher as well, the expert usually masters an analytic or semi-implicit, rather than synthetic, knowledge of his field of specialization. Translating his expertise into pedagogical discourse will require specific training on his part. We explain this unavoidable problem by the following: the specific knowledge of the expert can be considered as a body of answers to questions that no longer have to be asked. Jean Piaget, among others, defines cognitive operations as manipulations applied to memorized schemata.

> [Reflective abstraction] "consists in drawing from an inferior system of actions or operations certain elements which it reflects (...) on superior actions or operations, for one can only notice the processes of an earlier construction by means of reconstruction on another level"[11]

It is probably during this memorization process that the questions underlying the cognitive structures disappear, having completed their work of support.[12]

Thus, if one wants access to an expert's knowledge, one must then look for the questions to those answers. Reconstituting his expertisemeans walking the same road to discovery, rebuilding either the context in which occurred the discovery or the moment, namely apprenticeship, when scattered information coalesced into one body of knowledge. This is why interviewing the expert is so important.[13] During the interview, the knowledge engineer must simulate a question-and-answer routine similar to that which prevailed at the time when the expert made his initial discovery or apprenticeship. Spoken or written discourse being the instance where expertise is reconstituted, it is paramount that we take its expression into account. In order then for the implicit part of knowledge to be reduced to a minimum during the interview, reconstitution must be made through an explicit discursive process, that is "[an] intellectual operation achieved by a series of partial and successive basic operations".[14]

In a perspective broadened beyond the approach presented here, expertise reconstitution consists in implementing a set of genetic, cognitive and historical procedures. The genetic procedures work toward revealing the dynamics behind the elaboration of concepts (i.e. the translation from the concrete to the formal). The cognitive procedures concern the type and nature of operations behind the working of a given concept. Finally, the historical procedures stress the social dynamics (social command, paradigmatic trends, styles, etc.) restraining the production of one or a set of concepts.

3. Recording vs Taking Notes

Sooner or later, installing an expert system leads us to confront the many interpretations of the expertise involved. Should only one source of expertise be used, a set of interviews with an expert for example, we would only be pushing back this confrontation and make the integration of the various points of view only the more difficult afterwards. We then believe that expertise will be the better defined if it is approached from different venues, venue agreement thus pointing to stronger pockets of knowledge, venue disagreement to weaker ones. Moreover, to consider only those sources of human expertise is to neglect a great amount of available material. For the most part, the memory of an organization is of a textual nature and is stored in its records which consist of reports, memos, letters, etc.

Taking notes, whether it be during interviews with agents of the organization or while consulting archives, is a mode of information retrieval that turns out to be inadequate because of the subjective nature of the process and the problems it raises as for data validation. In interviews, selective transcription of notes or paraphrasing make it impossible to later distinguish the analyst's inferences from the expert's original answers. In the case of textual material, a vast corpus inhibits the perception of regularities whose occurrences are too distant from one another. In fact, loss of contact with the initial formulation makes it difficult to validate the cognitive structures retrieved and impossible to check their exhaustivity. The transcription of interviews on magnetic support may entail translation and revision costs, but it also saves time. Revision is also facilitated by the integration of technical lexicons, produced during the analytical process, within the word processing program used.

4. The Approach vs the Program

Our study of large-scale expertise transfer brought us to define an approach rather than a program. Among the many options offered by scientific reasoning, we thought it best, in order to deal with the inevitable losses due to the reduction process, to adopt a mainly inductive point of view. Induction[15] starts from specific, known objects: texts in this instance, as they appear to users. But if texts are those objects readily accessible for analysis, the information they contain however is unclear and not easily discernible from the discourse in which they are enmeshed. With the help of screening mechanisms peculiar to the text, ascent is attempted and the general cognitive structures of the corpus tracked down. This process is quite the opposite of deduction, which filters a preconceived notion of the final layout that will be scanned in a determinist fashion.

Most of the programs with which we are acquainted[16] project coding grids that, in our opinion, ignore the knowledge-making process of discourse to only consider sets of discrete units. There is direct bias when the analytic grid is determined in the procedure, indirect bias when the user is first called upon to construct a grid that willbe projected onto the text. However convenient from an operational standpoint, grid projection in both instances emphasizes a static description of knowledge.

We propose an interactive process where the heuristic dimension stands out: standardized manipulations devoid of content interpretation partially reveal expertise and direct explorations further on. This cyclical process consists of as many exploration (retrieval) / validation loops as are deemed necessary. Thus, not only do team members in charge of implementing an expert system validate its operations, they also control them. It could be said that our contribution is to make them aware of the existence and usefulness of software tools capable of enhancing the efficiency of the process. Our approach here does not take after the closed-off mill where confidence in tools reigns supreme, but rather the pocket calculator which through repeated, unrestricted and varied manipulation encourages user creativity.

5. Computer-Assisted Text Analysis

We believe the main obstacle to a optimal use of texts in expertise transfer lies both in an exclusive focus put on linguistic dimensions and in a lack of understanding of problems inherent to the text.[17] The words that compose it do not necessarily refer to reality through concepts, but may serve to recategorize them. Constant modifications in the referential structure of the text bring the reader to produce several inferences. This is why the text is said to function in a polysemous fashion and that the many interpretations it engenders never deplete it. Our mastery of natural languages is developing still and problems persist that have yet to be fully resolved within the scope of the sentence, i.e. anaphora,[18] incomplete wording, etc. Those problems diminish within the scope of larger segment because of the redundancy which caracterized texts. This is why we are looking for a theoritical framework.suited for textual reality.

In order to locate a list of potential concepts comprised in a vast corpus of texts, we try to integrate into our approach two methods of text analysis usually considered the opposite of one another: quantitative, based on statistical co-occurrences, and qualitative, based on an exhaustive description of syntactic structures. The search of concepts based on morpho-syntactic patterns increases exhaustivity, the scope of sources used and terminological precision. The survey is not conducted from any preconceived notions one might have about the data, nor from analogies drawn

with other fields, but from the data itself. In the other hand, the survey is not based on the syntactical correctness of the expressions because experience shows us that terms could be ruled out from a linguistic point of view but recognized by domain experts.

Furthermore, searches made from a morphological description on a great number of textual segments guarantee a constant, objective, reproducible and independent of problematics exposed in the text. Once a list of terms has been extracted and validated by domain experts, the concepts are reconstituted by means of classification of the syntagmatic configurations of the terms. By-products useful to the organization are generated throughout the procedure: lexicons allowing terminological standardization, textual databases available through the expert system or other means, and most importantly increased organizational awareness of the wealth stored in the texts it produces.

Within the general context then of expertise transfer, and since it is based on discourse analysis, our intervention fills a gap in methodology and thus allows the transformation of discourse into cognitive data. We intend to use the specific metalanguage of the text in order to isolate invariants hierarchically organized according to their recurrence. However, before showing the computer processes that carry out the transformation of data into structured objects submitted to the formal logic of expert systems, we deemed it necessary to define the "natural" logic of discourse as well as the objects of "schematization" it governs. But first we must elaborate on the subject of knowledge perception.

6. Unified Knowledge Perception

Most contributors to exertise transfer present a broken if not truncated view of knowledge. In an ontological perspective, discussions center around knowledge "primitives", metaknowledge, and on a knowledge of the knowledge of reality. In a typological perspective, knowledge is said to be procedural, analogic, probabilistic, etc., or else it is identified in comparison with objects to which it is related: substance, causality, temorality, etc. In order to locate allinformation within a text, we will not rely on a partially constructed, analytic vision of knowledge. We will refer instead to a more general framework that incorporates those positive dimensions of cognition already mentioned and the circumstantial aspects inherent to the conditions of its development, or in other words the discursiveness of its exposition.

This framework is inspired from Michel Foucault who defined knowledge as "[a] set of elements formed in a regular fashion through discursive practice".[19] Foucault's definition contains

two points of chief importance: the idea of regularity and that of discursive practice. The elements of a particular knowledge differ and vary from one to the next, yet they are also endowed with a cognitive and temporal stability that make it possible to hold symbolic activities such as (human or expert system) reasoning, argumentation, demonstration, etc. Stability here is not intrinsic to the objects, that are but mere constructions, but is assured through a practice of discourse directed in accordance with a given purpose. Expertise will thus be considered as a discursive structure "where among objects, types of enunciation, concepts and thematic choices some regularity can be outlined (an order, correlations, positions and functionings, transformations)".[20]

It clearly turns out that objects of knowledge are irreducible to the concepts of one domain of positive science.[21] Moreover, knowledge is multifold and its many expressions do not overlap entirely, for inany organization to the division of labor corresponds a division of knowledge, and knowledge gives way to other knowledge.[22] In other words, different simultaneous or consecutives discourses are held inside the organization by various agents or groups of agents using the same concepts. In a broader perspective than cognitive unit screening, one ought to take into account the social dimension of knowledge and consider, beyond those concepts that often are the ingredients of a heurism,[23] other factors such as common meaning, beliefs, cognitive processes peculiar to a "school" of knowledge, etc.

7. Natural Logic as a Reference Framework

Unlike logical formulations handled by expert systems, knowledge-conveying linguistic formulations are not subject to stringent rules of exposition. The very expression of similar objects can thus vary significantly. Partial formalizations can of course be drawn from the discourse itself: taxonomies, sector-based decision trees, algorithmic procedures, etc. But all the while, we overlook an important residual: the process of symbolic assimilation of reality, the transactions and negotiations from which concepts are elaborated.

Formal logic[24] usually disregards the nature of objects it manipulates. In a context of knowledge production, however, the objects handled are seldom nondescript and most certainly cannot be reduced to variables marked in a binary fashion (true or false). More often than not, these symbolic objects are identified by severalvariables whose values are not boolean but scalar, in that they belong to mereological classes. Just as the continuous stands opposite the discontinuous, so does the collective (or mereological) class set itself apart from the distributive (or set) class. We formally identify this opposition by the following: the concept of a distributive class is based on the relation "to be an element of" which is non-reflexive, asymmetric and intransitive; the other concept

of mereological class is based on the relation "to be part of" which itself is reflexive, symmetric and transitive. In short, a mereological class or mereonomy is a part-whole hierarchy. The human hand is a good example of such a hierarchy: each finger is not so much a part as it is an extension of the hand.

Prevailing within the mereological classes are those network relations between a whole and its parts, between subparts, etc.: the usual handling of symbolic objects does somehow appear to refer to other operational systems than those thematized by formal logic (inference through simple calculation of truth). This is why we steered away from a determinist formalism (predicate logic, etc.) and adopted instead a reference framework that takes into account the topological dynamics of knowledge enunciation. Indeed, the problem we sought to resolve was not "predictive" by nature but qualitative, spatial and temporal. Everyday reasoning, whether scientific or not, "conducted according to purpose, appears as a chain, a combination or a confrontation of statements and representations abiding by internal constraints that can be made more explicit".[25] We are then dealing with topological schemata rather than strictly determinist itineraries.

Circumstances surrounding everyday reasoning are most important, since signifying objects are not solely handled for demonstration purposes.[26] Four postulates mark this approach:

1) Each time Speaker A discourses, he proposes a schematization to Speaker B.
2) Speaker A's logical discursive activities are carried out in a determined situation.
3) The schematization Speaker A proposes to Speaker B is a function of A's purpose, and also of the representation A has of B, the relation he entertains with B and that which is in question, i.e. the topic (T).
4) This schematization consists of images of Speaker A, Speaker B and the topic T. It also bears the marks of its development.

In addition to deductive activities, manipulations may comprise validations, inductions, the development of hypotheses and analogies, etc. The description and definition of the terms of a particular knowledge is done in those of the natural language which stands in turn as its own metalanguage.[27] Furthermore, the meaning of phrases describing the terms of this knowledge is a function of operations and dealings inherent to the purpose of the instance of communication which governs statement validity and accuracy. We thus referred tonatural logic because it is concerned with the schematization processes brought about by speakers involved in discursive practices. These schematizations structure cognitive objects and position them within a field of knowledge, yet they always depend on specific circumstances, that is the social practices that determine the conditions of occurrence. In this perspective, schematizations are mostly characterized by their variety.[28]

8. The Objects of Schematization

Like objects of formal logic, objects of schematization are cores around which reasoning is elaborated. As we saw previously, they present one major difference: they can either constitute object or mereological classes (cf. supra); subparts of parts (e.g. "The justifications are often due to a curve problem, a deficient drainage problem, the road surface is too old, accident problems. All thas is the justification.");[29] archetypes of a particular class; or metonymies (e.g. "The Ministry of Transport knows the regulations as much as we do"[30] [here the whole encompasses its parts, i.e. the agents of the Ministry of Transport]). Within the framework of natural logic, the properties as well as any relation between objects of schematization are represented by predicates. One will thus encounter, in addition to those relations found in formal logic (implication, antithesis, equivalence, etc.), relations of object transformation, metafunctional relations (such as the introduction of a text or author), etc.

The anchorage procedure is the instance through which semantic cognitive units, once fixed in noun or verb forms, take place in the schematization process. Because they refer to reality, these noun and verb forms are to be taken as substantives. Noun anchorages embody within discourse mereological object classes. A concept for example such as "project" does not have any "meaning" in itself; it only gains meaning from elements (ingredients) which outline its limits (e.g. "The project under consideration consists of repairs to the liquid waste emission for the paper mill C...).[31] Verb anchorages provide elements of object dynamics - properties and relations (e.g. The project aims to improve wildfowl production for the Black lake swamp which covers an area of 47 hectares).[32]

Let us recall that objects of schematization are recursive, always summoned and reworded by speakers throughout the discursive process: several nominal substantives can successively refer for example to the same object (synonymy). Let us also ponder the fact that to a given noun phrase corresponds a specific way of structuring reference to the object. Consider the following question and answer sequence:

"How do you apprehend the repercussions analysis?"

"Some of the residences are affected, noise problems (...) A few lakes are affected.

(...) A golf course is affected."[33]

The concept here of "répercussion" is analyzed from many perspectives. This exemplifies how the "meaning" of a word or term is the manner in which signification corresponds to extralinguistic reality. Let us illustrate our point with the following example, drawn from arithmetics. The

statements 2 + 2 and 1 + 3 are equivalent in that 2 + 2 = 1 + 3 = 4, but only in terms of the result obtained are these statements thus equivalent; the manner in which the result was obtained is totally different. We can reasonably assume that the act of referencing, whether it focuses on arithmetic or discursive elements, is similar in design.

9. Morphological Discourse Analysis

Objects of schematization are not linked like the arguments of a formal demonstration; they are shaped following set courses we must reconstitute and contextual relations we must pinpoint. This terminology refers to the theoretical core that forms itself around the topology. The property the objects of schematization have of being constructed, transformed and modified by speakers is revealed in the variety of anchorages comprising these objects and the diversity of predicates. In this case, relations between objects can only be apprehended through the category of space. This is why we speak of a topology, where relations of proximity and distance mark the many parts of a discourse or text: the introduction has to stand out against the conclusion and the body of the text; the problematic has to differ from the analysis; the arguments have to outline theattitude of the subject in relation with objects at stake in the discourse or text; the list goes on.

Morphological discourse analysis,[34] which inspired our own approach, is based for the most part on the hypothesis that statements of a discourse appear as regular-shaped object-core forms and networks. Analyzing discourse morphology amounts to constructing a general model of the text by listing the objects of schematization found in its syntactic strata and by reconstituting, beyond the strict boundaries of sentences, the semantic itineraries followed by these objects. Morphological discourse analysis capitalizes on the distinctive feature of natural languages of being their own metalanguages, that is to say that they can both depict reality and the representation of reality. This validates a reading by retrieval and sampling of textual segments (the process is known in technical terms as "thematization by specification") describing the major stakes of the discourse or text. Structured together, these segments form a new text that presents itself as a result of interpretive practice.

Sequentiality[35] is that process by which objects are put into discourse and later reconstituted through reading (a particular discursive process in itself). Textual sequence construction is carried out in this process along a nominal and verbal axis. In the first instance, the text is apprehended from the relations woven by noun forms, e.g. the systematic repeating of a semantic category by means of various nominal or pronominal (anaphoric) phrases. In thesecond instance, verb forms and gerunds (nouns derived from verbs) found a logic of action by steering the courses taken by

statement subjects. Some verb forms are thus used to mark the opposition between the continuous and the discontinuous, between potentiality and actuality, etc. In this perspective, natural logic guides our study of object structuring; grammar (semantics + syntax) helps us in singling out the material regularities of the language in which these objects are represented.

10. Our Analysis

A first step consists in locating concepts or cognitive units of the field of expertise and composing a dictionary that will later be transposed into frames. This mode of knowledge representation is prefered over predicates since it offers modularity, flexibility and readability and above all because it outlines the maximal extension of each concept. These primitives will later be needed to define the scope and operators of the problem at hand in terms of inference rules (premise conditions as well as inferences) and facts (the data of the problem).

In order to establish this dictionary, we analyze nominal substantives in terms of the configuration of words (also called ingredients) that are related to them. For a substantive such as "projet" for example, we may have structures such as "l'assujettissement d'un projet", "la pertinence d'un projet", etc. This analysis is justified in that the effect of reference to reality (the concepts in this particular case) in a given discourse depends on noun forms that consolidate other noun forms into object categories. Referential markings[36] are then obtained from statement structures and linear transformations that generate textual dynamics. These markings are linguistically identifiable on account of discourse strategies[37] that confer textual control on some noun forms. To illustrate our point, we submit the following structures:

.Finite description + anaphoric pronoun sequence:

Object (x) (...) "it","it", etc.

.PP + deictic sequence | anaphoric substantive:

"the project is not completed (...) the project is behind schedule (...) this is a case of reject (...)"[38]

.PP + nominalization sequence:

PP1, PP2, PP3 ... "The request is complete"[39]

.Impersonal construction + (NP | PP) + "que" + PP*:

"There are cirteria one has to respect: a, b, c (...)"[40]

This shows how different statement structures can be arranged in a linear fashion according to reasoning purposes. A "natural logic" approach stresses the relations of transition between statements of a discourse or text.

The three steps in composing a dictionary of valued objects are:

i) Establishing a list of concepts
    Morphological categorization
    Phrase locking
    Purging the word list
    Drawing up synonyms
ii) Detecting features
iii) Detecting values

The writing of inference rules will come at a later stage when we analyze determination, that is the type of relations between objects and the operations likely to be applied on them. A lexicon of operations will then be set up from transitive verbs which describe transformations undergone by valued objects; in each operation,arguments are linked in terms of valued objects. Verbs of action and state will also be analyzed in terms of modulation (active, passive, necessary, elective, etc.), localization and temporality.

Later on, those segments which are significant will be outlined according to content uniformity and categorized following this tentative grid: definitions, descriptions, rules, axioms and strategies, the latter segments indicating how other segments should be rearranged in order to follow the order of operations needed to solve the problem. Further on, we will pay close attention to enunciation itself, that is the forms of inscription of determination by the subject in a (social) context of communication, and the variations in objects of schematization (e.g. "submit a request", "propose a project", etc.).[41] Enunciation concerns us because it governs the dynamic aspects of the production of schematizations. Following steps in his reasoning, the expert can switch from statement to action, from the positive to the probable, and so on. Statements will then be structured, in a more or less complex fashion, into significant textual segments (1,2,...,n-statement configurations). These appear as conjunctions, concessives, restrictives, transitives, etc. (e.g. "the project application should follow the steps a,b and c"),[42] and as hypotheses and consequences. This analysis will enable us to show the transitional states of the problem and the control strategies that will lead to its resolution.

11. Establishing a List of Concepts

This first step consists in going from the lexicon of the analyzed corpus, that is the body of words that constitutes the corpus and their frequency, to the discursive concepts or terms, understood as cores of noun anchorages. The process breaks down into the several following operations.

We first proceed with the morphological labelling of the lexicon. We then lock on terminological phrases, i.e. those terms that transcend the lexical unit. A great number of technical terms are composed of more than just one word, each having its own meaning when taken separately (e.g. "word processing"). Their construction around a nominal head seems to depend on morpho-syntactic patterns. This observation enables us to retrieve textual segments (concordances) from the co-occurrence of morphological patterns.[43] Consider for example these structural patterns we encountered:

.[Noun + "de" + Noun]:

"avis de projet", "centre de documentation", "chargé de projet"

(project application, documentation center, project evaluator)

.[Noun + Adjective]: because of the general nature of this pattern, it was extended to [Noun + Adjective + Adjective]:

"réunion générale annuelle" (annual general reunion)

Not all segments obtained are terminological phrases; an expert system must therefore purge the list of lexical items. This procedure rests on value judgments and requires rigor and consistency: a word chain that appears by itself at a higher frequency rate than that of the nominal head of the phrase is generally a sure indication. There are, however, no absolute selection criteria. Rather, these criteria relate to the field of expertise to be covered and the expected configuration of cognitive units. On the one hand, it would be erroneous to only consider the chain "repercussion analysis" because we also find "impact analysis", which thus allows the construction of the cognitive unit analysis -» type (repercussion/impact)]. On the other hand, we could only consider the chain "repercussion analysis" if the cognitive unit anticipated was more general: [step -» type (repercussion analysis/impact analysis)]. In all, we must be careful not to link concepts with their values, should a feature be implicit, on the sole basis of a number of co-occurrences. Terminological phrases should not exist per se but in relation to their intended use. These phrases will bear a nominal label.

Once the terminological phrases have been locked, the lexicon of the corpus is restricted to nominals then reduced manually by an expert in order to keep only those concepts related to the field of expertise. For example, SAGÉE's lexicon of concepts will comprise such phrases as "project application", "dredging", "sediment", etc.[44]

Equivalent concepts found in the lexicon are then brought back manually to a common, canonical form: in this perspective, "noise problem" and "sound pollution" are considered

synonyms. Still, synonymy must first be validated by the expert. Should these concepts relate to two different, consecutive stages in the course of a process, there will be no equivalence. This operation enables the analyst to reduce the number of concepts and, while scanning the corpus, get hold of the maximum number of occurrences of a term from its synonyms.

The next chapters describe the change from concepts to valued objects. Their characteristics and number account for the richness in inference of the expert system. In order to properly use the textual material at hand, noun phrases whose heads correspond to the concepts of the lexicon must be linked to the object-feature-value structure.

12. Linking Features

The maximal extension of each concept/object is ascertained by superimposing all their contexts of occurrence (concordances) in the corpus. Many features can be detected using simple structures such as [Feature + Preposition + Concept]. A concept such as "quay", for example, will be found in the following segments:

"The total quay length (...)"
"The quay location (...)"
"If the quay is located (...)"
"The quay width (...)"[45]

Within the perspective of morphological discourse analysis, we work toward finding more complex patterns that will enable us to locate configurations that might escape the concordances. During this analysis, informations concerning the inscription of concepts in a hierarchical structure should be located around such partitive phrases as "a kind of", "a part of", etc. These links must be recorded with care and tagged to related concepts in the dictionary.

13. Detecting Values

Once the nominal configurations that give the objects their features have been detected, the analysis is extended to the examination of adjectives found in the contexts drawn up. Among the different adjectival forms looked for, there are quantifiers (numerals, cardinals and ordinals), forms that identify degrees in a scale (e.g. cold, mild, warm, lukewarm, burning, boiling, etc.),[46] and mass partitives (e.g. half, three quarter, etc.).[47] These forms are an expression of the argumentative scales that virtually position all other possible qualitative or quantitative values.[48]

The next step, once the analysis of nominal substantives is completed, is to constitute a dictionary of concepts/objects extensively described in terms of their characteristics (their features) for which the admissibility of the value is specified in terms of scale or constraint. This dictionary may be used to such ends as generating the cognitive structures of an expert system. Its validity has yet to be tested on several bodies of texts and an evaluation of the results has yet to be done.

14. Conclusion

In this paper, we proposed an approach inspired from a discourse analysis based on an unified and "natural" perception of knowledge. Here, the detection of concepts describing the scope of a problem meets criteria of consistency, objectivity, reproductiveness and independence as to the problematics defined in the texts. The goal of this interactive approach, which can be suitably applied in real time on a great corpus, is to promote user creativity.

Faced with an always growing mass of texts produced by organizations which exceeds by far reading opportunities, we advocate the use of computer packages for the processing of knowledge consequent to that of the texts themselves in order to enhance their considerable source of expertise.

---

[1] The authors can be reached at the following address :

Centre d'Analyse de Textes par Ordinateur
Université du Québec à Montréal
P.O. Box 8888, Station A
Montreal, QC
H3C 3P8 Canada

Tel.: (514) 987-8256
Fax: (514) 987-4567
E-mail: PAQUIN@ATOCI.UQAM.CA

[2] We want to mark the important contribution of Alain Lecomte (GRAD, Grenoble) and Jean-Marie Marandin (LISH) to the field of discourse analysis and especially to the development of the hypotheses we discuss in the course of this article.

[3] G.A. Kelly, *The Psychology of Personal Constructs*, New York: Norton, 1955.

[4] K.A. Ericcson & H.A. Simon, *Protocol Analysis: verbal reports as data*, Cambridge (Mass.): MIT Press, 1984.

[5] S. Grosberg, *Neural Network and Natural Intelligence*, Cambridge (Mass.): MIT Press, 1988.

[6] Cf. on the subject: W. Frey, U. Reyle & C. Rohrer, "Automatic Construction of a Knowledge Base by Analysing Texts in Natural Language", *International Joint Conference on Artificial Intelligence*, 1983, pp. 727-729; and T. Nishida, A. Kosaka & S. Doshita, "Towards Knowledge Acquisition from Natural Language Documents - Automatic Model Construction from Hardware Manuals", *International Joint Conference on Artificial Intelligence*, 1983, pp. 482-486. One system includes this functionality for German: cf. J. Diederich, I. Ruhmann & M. May, "KRITON: a knowledge-acquisition tool for expert systems", *International Journal of Man Machine Studies*, 26 (1987), pp. 29-40.

[7] The Système d'Aide à la Gestion des Évaluations Environnementales (SAGÉE) has already been the object of a paper: cf. *Actes du premier colloque québécois en informatique cognitive des organisations*, 1987, Section 2, pp. 21-28.

[8] The present size of the corpus is estimated at 3 million words (15 meg).

[9] Among others : *Development of a computer-assisted content analysis system* known as "Système d'Analyse de Contenu Assistée par Ordinateur (SACAO)" In this research, headed by Jules Duchastel and financed by the Fonds FCAR of Quebec within its "Actions spontanées", both the specifications and the methodology were drawn; *Worbench for knowledge engineering and textual data analysis* known as "Atelier cognitif et textuel (ACTE)". This software integration program of two applications, one for textual data analysis (SATO) and the other for expert system building (D_expert) will enable the building of complex textual data analyser by its users without an intensive computer science training. This project is financed by a governmental agencies consortium where pilot projects have taken place earlier. Lately a research group was formed to explore and experiment the parallel computing paradigm in relation with textual data analysis of very large body of texts (over one gigabyte). This project take place in the frame of a R.&D. tax exemption program for investors.

[10] A dictionary of concepts does appear necessary in many approaches to expertise transfer: cf. A. Hart, *Knowledge Acquisition for Expert Systems*, New York: McGraw-Hill, 1986, pp. 65-66.

[11] [L'abstraction réflexive] consiste à tirer d'un système d'actions ou d'opérations de niveau inférieur certains caractères dont elle assure la réflexion (...) sur des actions ou opérations de niveau supérieur, car il n'est possible de prendre conscience des processus d'une construction antérieure qu'au moyen d'une reconstruction sur un nouveau plan" : cf. his *Études d'épistémologie génétique*, Paris: Presses universitaires de France, 1961, XIV, p. 203.

[12] Michel Meyer, *Découverte et justification en science*, Paris: Éditions Klincksieck, 1979: his Chapitre IX, "La Conception problématologique de la science", pp. 289-353.

[13] "A systems analyst can fact-find by questionnaire, by sampling records or by observing people at work, but no analysis is complete without face-to-face discussions with users": A. Hart, Knowledge *Acquisition for Expert Systems*, p. 49.

[14] "[une] opération intellectuelle qui s'effectue par une suite d'opérations élémentaires partielles et successives": André Lalande, *Vocabulaire technique et critique de la philosophie*, Paris: Presses universitaires de France, 1976, p. 237.

[15] This definition of induction is directly inspired from Aristotle (*Physicorum* I, 184a).

[16] Among others: J.S. Bennett, "ROGET: a knowledge-based system for acquiring the conceptual structure of a diagnostic expert system", *Journal of Automated Reasoning*, 1 (1985), pp. 49-74; J.H. Boose, "A Knowledge Acquisition Program for Expert System Based on Personal Construct Psychology", *International Journal of Man Machine Studies*, 23 (1985), pp. 495-525; J.M. Bradshaw, "Expertise Transfer and Complex Problems: using AQUINAS as a knowledge-acquisition workbench for knowledge-based systems", *International Journal of Man Machine Studies*, 26 (1987), pp. 3-28; L. Eshelman, D. Ehret, J. McDermott & M. Tang, "MOLE: a tenacious knowledge-acquisition tool", *International Journal of Man Machine Studies*, 26 (1987), pp. 41-54; G. Klinger, J. Bentolina, S. Genetet, M. Grimes & J. McDermott, "KNACK: report-driven knowledge acquisition", *International Journal of Man Machine Studies*, 26 (1987), pp. 65-79; and N. Aussenac & B. Michez, "M.A.C.A.O.: application d'un modèle psychologique à la réalisation d'un outil d'aide à l'acquisition des connaissances", *Actes du colloque «Représentation du réel et informatisation»*, Saint-Étienne (France), (26-27 May) 1988.

[17] Discourse analysis is the field which focuses on these problems. For an overall view of the subject, cf. D. Maingueneau, *Initiation aux méthodes de l'analyse du discours*, Paris: Hachette, 1976.

[18] Anaphora (from the Greek meaning: reference, recall, recourse) describes the characteristic some words (usually pronouns) have of linking (referencing) words or terms to other elements stated earlier, e.g. "he" in "Peter is eating an apple, he likes fruits". Here, the pronoun "he" refers to "Peter". There are many types of anaphoric relations: pronouns used as substitutes, terms that consolidate sequences of statements (e.g. from our corpus: "le dragage, l'excavation, le pavage ... ces travaux seront requis"), etc.

[19] "[un] ensemble d'éléments, formés de manière régulière ar une pratique discursive": Michel Foucault, *L'Archéologie du savoir*, Paris: Gallimard, 1969, p. 238.

[20] "où entre les objets, les types d'énonciation, les concepts, les choix thématiques on pourrait définir une régularité (un ordre, des corrélations, des positions et des fonctionnements, des transformations)": Michel Foucault, *L'Archéologie du savoir*, p. 53.

[21] Cf. Pierre Bourdieu, "La Spécificité du champ scientifique et les conditions sociales du progrès de la raison", *Sociologie et société*, 7, 1: Science et structure sociale, pp. 91-116.

[22] J.P. Poitou, "The Expert and the System", a paper submitted by the author in May 1987.

[23] G. Polya, *How to Solve It. A New Aspect of Mathematical Method*, Princeton (NJ): Princeton University Press, 1973.

[24] For a classification, cf. Susan Haack, *Philosophy of Logics*, Great Britain: Cambridge University Press, 1978, p. 276.

[25] "se présente comme un enchaînement, une combinaison ou une confrontation d'énoncés ou de représentations, respectant des contraintes internes susceptibles d'être explicitées, conduit en fonction d'un but": Pierre Oléron, *Le Raisonnemnt*, Paris: Presses universitaires de France (Que sais-je? #1671), 1977, p. 9.

[26] M.-J. Borel, J.-B. Grize & D. Miéville, "Essai de logique naturelle", *Sciences pour la communication* (4), Berne: Éditions Peter Lang SA, 1983, p. 99.

[27] Antoine Culioli, "La Formalisation en linguistique", in *Considérations théoriques à propos du traitement formel du langage*, A. Culioli, C. Fuchs & M. Pêcheux, Paris: Dunod (Documents de linguistique quantitative #7), 1970, pp. 1-13.

[28] M.-J. Borel, J.-B. Grize & D. Miéville, "Essai de logique naturelle", pp. 99-146, 241.

[29] In French: "Les justifications c'est souvent à cause d'un problème de courbe, un problème de mauvais drainage de la route, la chaussée est toute désuette, des problèmes d'accident. Tout ça c'est la justification"

All examples are drawn from interviews conducted within the SAGÉE project: we underline. Since this is retranscribed spoken discourse, the syntax is less formal.

[30] In French: "Le ministère des Transports connaît les règlements autant que nous autres"

[31] In French: "Le projet à l'étude consiste en la réfection de l'émission d'eaux usées de l'usine de pâtes et papier C".

[32] In French: "Le projet a pour objectif d'améliorer la production de sauvagine du marais Lac Noir (comté de l'Islet) qui a une superficie de 47 hectares"

[33] In French: "Au niveau des répercussions analysées, comment appréhender quand vous analysez ce point?"

"Il y a des résidences d'affectées, des problèmes de bruit. (...) Il y a quelques lacs affectés. (…) un terrain de golf d'affecté."

[34] A. Lecomte & J.-M. Marandin, "Analyse de discours et morphologie discursive", working paper, 1984, 67 p.

[35] A. Lecomte, "Espace des séquences: approche topologique et informatique de la séquence", *Langages*, 83 (1986): Analyse de discours: nouveaux parcours", D. Maldidier et al., p. 93.

[36] We draw here on a paper presented by Alain Lecomte at the ACP "ADELA" seminar on 27 January 1983, "Algorithmes de la séquence", 24 p.

[37] Namely the choice of themes and the nominal instancing of semantic primitives; Michel Foucault, *L'Archéologie du savoir*, pp. 85-93.

[38] In French: "le projet n'est pas complet (...) le projet retarde (...) c'est un cas de rejet (...)"

[39] In French: "la demande est complète"

[40] In French: "Il y a des critères qu'il faut respecter: a, b, c (...)"

[41] In French: "soumettre une demande", "proposer un projet", etc.

[42] In French: "la demande de projet doit passer par l'étape a et b et c"

[43] Conceived by F. Daoust, SATO is able to lock on concordances found in the text and then establish a frequency lexicon; it can also help to draw phrase listings that can be projected onto other texts.

[44] In French "avis de projet", "dragage", "sédiment", etc.

[45] In French "La longueur totale du quai (...)", "L'emplacement du quai (...)", "Si le quai est situé à (...)", "La largeur du quai (...)"

[46] In French "froid, tiède, chaud, brûlant, bouillant, etc."

[47] In French "demi, trois-quart, etc."

[48] Cf. Oswald Ducrot, *Les Échelles argumentatives*, Paris: Minuit, 1980.